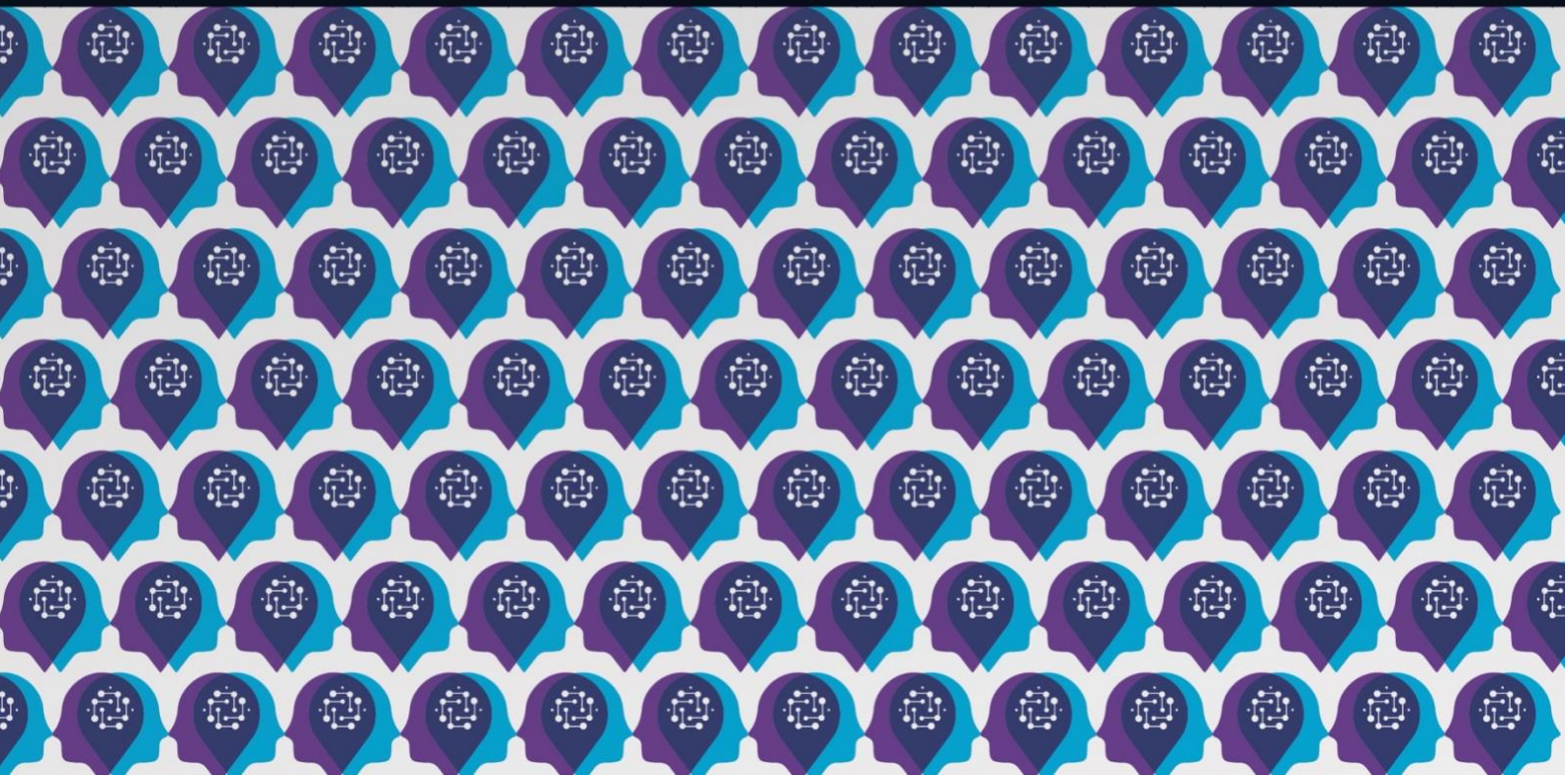
AI4Debunk

D12.2 Resilience mechanisms triggered by the tool

January 2025

Funded by
the European Union



Grant Agreement No.: 101135757
Call: HORIZON-CL4-2023-HUMAN-01-CNECT
Topic: HORIZON-CL4-2023-HUMAN-01-05
Type of action: HORIZON Innovation Actions

D12.2 RESILIENCE MECHANISMS TRIGGERED BY THE TOOL

Guidelines and Recommendations for Tool Developers

Project Acronym	AI4Debunk
Project Number	101135757
Project Full Title	Participative Assistive AI-powered Tools for Supporting Trustworthy Online Activity of Citizens and Debunking Disinformation
Work package	WP 12
Task	Task 12.2
Due date	31/03/2025
Submission date	31/03/2025
Deliverable lead	Pilot4dev
Version	V1.0
Authors	Pascaline Gaborit, Vishnu Rao, Joen Martinsen
Contributors	-
Reviewers	Matei Mancas (University of Mons - UMONS)
Abstract	This deliverable proposes the first general guidelines for the AI4DEBUNK tools' developers. It addresses the points of User Friendliness, Social Media, Inauthentic Coordinated Behavior, Ethics, Integration of Multi-Languages, 'explaining Fakeness', and 'Stakeholders integration'. The Annex II on the meeting with the beta testing group is also added.
Keywords :	Guidelines, AI tool, Inauthentic Coordinated Behaviour, Improvement, Beta testing

Funded by the European Union

DOCUMENT DISSEMINATION LEVEL

Dissemination level	
X	PU – Public
	SEN – Sensitive

DOCUMENT REVISION HISTORY

Version	Date	Description of change	List of contributor(s)
0.1	20/11/2024	First version	P4D
0.2	18/12/2024	Second version after quality assessment review	P4D
0.3	19/12/2024	Internal Quality Assessment Review	UMONS
0.4	23/12/2024	Project Coordinator Review	UL
1.0	30/01/2025	Final version ready for submission	P4D

STATEMENT ON MAINSTREAMING GENDER

The AI4Debunk consortium is committed to including gender and intersectionality as a transversal aspect in the project's activities. In line with EU guidelines and objectives, all partners – including the authors of this deliverable – recognise the importance of advancing gender analysis and sex-disaggregated data collection in the development of scientific research. Therefore, we commit to paying particular attention to including, monitoring, and periodically evaluating the participation of different genders in all activities developed within the project, including workshops, webinars and events but also surveys, interviews and research, in general. While applying a non-binary approach to data collection and promoting the participation of all genders in the activities, the partners will periodically reflect and inform about the limitations of their approach. Through an iterative learning process, they commit to plan and implement strategies that maximise the inclusion of more and more intersectional perspectives in their activities.

Funded by the European Union

DISCLAIMER

The AI4Debunk project has received funding from the European Union's Horizon Europe Programme under the Grant Agreement No. 101135757.

COPYRIGHT NOTICE

How to cite this report: **AI4Debunk (2025): Gaborit P., Rao. V., Martinsen J.Deliverable 12.2. Resilience Mechanisms Triggered by the Tools.**

The AI4Debunk consortium is the following:

Participant number	Participant organisation name	Short name	Country
1	LATVIJAS UNIVERSITATE	UL	LV
2	FREE MEDIA BULGARIA	EURACTIV	BE
3	PILOT4DEV	P4D	BE
4	INTERNEWS UKRAINE	IUA	UA
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR-IRPPS	IT
6	UNIVERSITA DEGLI STUDI DI FIRENZE	MICC/UNIFI	IT
6.1	CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI	CNIT	IT
7	BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION	BSC	ES
8	DOTSOFT OLOKLIROMENES EFARMOGES DIADIKTIOY KAI VASEON DEDOMENON AE	DOTSOFT	EL
9	UNIVERSITE DE MONS	UMONS	BE
10	NATIONAL UNIVERSITY OF IRELAND GALWAY	NUIG	IE
11	STICHTING HOGESCHOOL UTRECHT	HU	NL
12	STICHTING INNOVATIVE POWER	IP	NL
13	F6S NETWORK IRELAND LIMITED	F6S	IE

Funded by the European Union

TABLE OF CONTENTS

Funded by the European Union

Funded by the European Union

ABBREVIATIONS

AI	Artificial Intelligence
WP	Work Package
EC	European Commission
DSA	Digital Services Act
MAU	Monthly Active Users
VLOP	Very Large Online Platform

Funded by the European Union

EXECUTIVE SUMMARY

In the rapidly evolving landscape of online information, disinformation and misinformation pose significant threats to public discourse, political stability, and individual decision-making. This document outlines a comprehensive approach to enhancing the development and effectiveness of AI tools designed to detect and combat disinformation based on the resilience mechanisms and the first results of the beta testing group (Annex). The guidelines emphasize the importance of key factors such as user-friendliness, inclusivity, ethical considerations, and stakeholder integration. Critical areas like tackling coordinated inauthentic behavior, ensuring transparency in disinformation scoring, and staying updated on emerging disinformation tactics are thoroughly addressed. By continuously incorporating user feedback, regularly updating algorithms, and collaborating with experts, these tools can remain relevant, accurate, and responsive to new challenges in detection of disinformation. This holistic approach ensures that AI tools are not only effective in identifying false information but are also adaptable to evolving threats and user needs.

Funded by
the European Union

1 INTRODUCTION

The rise of social media and digital communication platforms has amplified the spread of disinformation and misinformation, creating challenges for individuals, organizations, and governments seeking to maintain an informed public. Disinformation campaigns—often coordinated and intentional—have the power to influence elections, manipulate public opinion, and undermine trust in institutions. Meanwhile, misinformation, though unintentional, can spread just as rapidly, causing confusion and potential harm.
In response, AI-powered tools have emerged as a critical means of detecting, flagging, and mitigating the impact of false information online. However, the development and maintenance of these tools require a multi-faceted approach that considers a wide range of factors, from user experience to algorithmic accuracy and ethical standards. This document provides a set of guidelines and recommendations aimed at enhancing the effectiveness of AI disinformation detection tools.
This project also foresees the set-up of a beta testing group to support and test the tool's development. The guidelines for the beta testing meetings are also added to this report. By incorporating insights through this first recommendations' report, and later on from journalists, researchers, policymakers, and other key stakeholders, these tools can be designed to meet the needs of diverse users. The continuous improvement of these tools—through regular updates, user feedback mechanisms, and collaboration with experts—ensures that they remain capable of addressing the ever-evolving landscape of disinformation. The following sections outline the best practices and recommendations for achieving these goals.

These recommendations are general, and not yet focused on the technical results of the AI4DEBUNK project.

Funded by the European Union

2 USER FRIENDLINESS

Creating a user-friendly experience is paramount to ensuring broad adoption and effective use of AI tools designed to combat disinformation. A user-friendly interface will not only help non-technical users navigate the tool but also allow more advanced users to benefit from its deeper functionalities. It is crucial that the tool is easy to operate, understand, and customize, catering to diverse user needs. Below is an in-depth look at the essential elements contributing to user-friendliness:

2.1 A) DESIGN INTUITIVE INTERFACES

- Ease of Use: The tool's interface should be intuitive, allowing users to navigate without needing advanced technical knowledge. This can be achieved by employing a clean layout with simple navigation, logical menu structures, and clear labels that guide users to different functions. For instance, icons and buttons should be clearly labeled to represent their respective actions (e.g., "Analyze," "Flag," "Report"), reducing any ambiguity in usage. For non-technical users, minimizing jargon and providing simple instructions on the tool's main dashboard can make it more accessible ,
- Streamlined Workflow: Simplifying the user workflow is critical. Tasks such as uploading content for analysis, interpreting results, and flagging suspicious material should be straightforward. Each step in the workflow should be visually and textually guided, helping users progress without feeling overwhelmed by complex options or unnecessary features. For example, a one-click feature to analyze content or a clear progress bar can help users track their tasks easily.
- Visual Aids: Effective visualizations—such as graphs, charts, and scorecards—can significantly aid in user comprehension of complex datasets. For example, displaying disinformation scores with color-coded risk levels (e.g., green for safe, red for potentially harmful) will help users quickly understand the seriousness of flagged content without diving deep into technical explanations. Graphs showing trends in how content is shared across platforms or how quickly it spreads can also provide clear insights. These visual aids make it easier to grasp the broader impact of the analyzed content.
- On-Screen Tooltips: Tooltips provide users with helpful information on-the-fly. These brief pop-up text boxes appear when a user hovers over a button or feature, explaining its function. For example, hovering over a "Flag Content" button could display a tooltip like "Click here to mark content for further analysis based on suspicious patterns." Tooltips reduce the need for users to consult external help resources, offering contextual support during the user journey. These tooltips depend on the possibility to have a mouse. Basically the interface should also be able to work on different kind of interfaces : big screens with mouse interaction, big touch-screens, small touch-screens. In the 2 latter cases the idea here cannot be applied. Also, the possibility to provide an interface adapted to different devices is important.

Funded by the European Union

2.2 B) CUSTOMIZABLE SETTINGS

- Personalization: Users have different needs and preferences. Customizable settings allow the tool to adapt to various workflows, whether it's a casual user seeking simple results or an advanced analyst looking for in-depth insights. Users should be able to adjust the level of analysis, set preferences for specific types of disinformation (e.g., political vs. health-related disinformation), or configure how they view data (e.g., summary view vs. detailed view.
- Flexible Alerts and Notifications: Some users may want real-time notifications of flagged content, while others may prefer daily or weekly summaries. The tool should provide flexible options for how often and in what form users receive updates—either through push notifications, email alerts, or in-app updates. For example, a user who monitors news media might want immediate alerts for disinformation related to breaking stories, while a research analyst might prefer a cumulative weekly report of all flagged content. In the case of AI4DEBUNK, the user will send a news/post to the application to check it, and not receive further notification. The project may however recommend registration to factcheckers if the users are interested to receive further notifications.
- Scalable Complexity: The tool should offer varying levels of complexity in its interface. Basic users can benefit from simplified reports, where data is aggregated and key insights are highlighted. Advanced users may require access to more granular data (e.g., a deep dive into a content's propagation network or detailed engagement metrics). For example, a researcher could toggle between "Basic" and "Advanced" modes, switching from high-level insights to detailed network analysis of how disinformation spreads across social media platforms.

2.3 C) ACCESSIBLE USER GUIDE

- Comprehensive Documentation: A detailed user guide is essential for ensuring that users understand how to utilize the tool's full range of features. The guide should be clearly structured, beginning with basic functionalities and moving to more advanced capabilities. It should include real-world examples that explain how users can apply each feature. For example, a section on "Understanding Disinformation Scores" could walk users through various scoring scenarios, explaining how the tool calculates a score based on factors like source reliability and network behaviour.
- On-Demand Help: Users should have access to on-demand help directly within the tool. This could include a built-in FAQ section, a help button for immediate assistance, or even contextual help that guides users as they navigate the tool. For instance, when users hover over a complex feature, a short explanation could pop up to describe its purpose. This ensures that users can find answers without needing to leave the platform to search for help.

Funded by the European Union

- Multimedia Learning Aids: Offering a variety of learning materials can accommodate different user preferences. Video tutorials can visually walk users through important features, providing step-by-step instructions that are easier to follow than text alone. Interactive walkthroughs can allow users to explore features in a simulated environment, learning by doing rather than reading. Diagrams and flowcharts could be used to simplify explanations of complex processes, like how disinformation is tracked or flagged.

2.4 CONCLUSION

Designing a user-friendly tool requires careful attention to how users interact with the interface, how much control they have over customization, and the ease with which they can learn to use it. By providing a clean, intuitive interface with clear visual aids and on-screen tooltips, developers can ensure that the tool is accessible to a broad audience. Offering customizable settings and scalable complexity ensures that users from varying technical backgrounds can adjust the tool to their specific needs. Finally, a comprehensive and accessible user guide equipped with multimedia aids ensures that all users, regardless of their learning style, can quickly become proficient with the tool. By prioritizing user-friendliness, developers will significantly improve the tool's accessibility, effectiveness, and adoption rate.

Funded by the European Union

3 ABOUT FIGURES, TABLES & REFERENCES

Inclusivity is a cornerstone of designing AI tools for disinformation detection, as these tools need to be accessible and effective for a wide array of users from different backgrounds, cultures, and abilities. Ensuring inclusivity means supporting multiple languages, accommodating users with varying levels of technical proficiency and physical abilities, and being sensitive to the diverse cultural contexts in which the tool will be used. An inclusive design guarantees that the tool is not only universally accessible but also adaptable to the nuances of different regions, linguistic needs, and socio-economic backgrounds. Below is a more detailed breakdown of how inclusivity can be implemented in AI disinformation tools.

3.1 SUPPORT MULTIPLE LANGUAGES

- Comprehensive Multilingual Support: To be truly inclusive, the tool must be capable of supporting multiple languages beyond simple translations of its interface. It is essential that the tool can analyze content in various languages, applying language-specific models that can accurately process linguistic features such as syntax, grammar, idioms, and colloquialisms. Disinformation often uses subtle language manipulations or regional expressions that generic models might miss. Therefore, the AI needs to be trained on language-specific datasets that reflect the unique structures and cultural references of different languages.
- Region-Specific Customization: The tool should offer regional customization options, allowing users to adjust settings or algorithms based on the specific disinformation challenges in their locale. For example, in regions experiencing political unrest, disinformation may focus on political figures or institutions, while in others, it might revolve around public health issues. Offering such customization ensures that the tool remains relevant and responsive to local disinformation patterns, helping users detect the most pertinent content.
- Cross-Language Insights: Advanced features could include cross-language comparisons, allowing researchers and users to track how a disinformation narrative spreads across different regions and languages. For example, a health-related hoax originating in one country might evolve and propagate differently in another country, even in a different language. The ability to monitor such cross-language disinformation patterns enables a more global understanding of how fake news is disseminated.

3.2 ACCESSIBILITY FOR ALL USERS

- Inclusive Design for Disabilities: To cater to a wide range of users, the tool must include features that support those with visual, auditory, or motor disabilities. This can be achieved by ensuring the tool is screen-reader compatible for users with visual impairments, and by incorporating

Funded by the European Union

keyboard navigation for users who may have difficulty using a mouse or touchscreen. This allows users with disabilities to navigate the tool's features with ease and engage with its outputs fully.

- Adjustable Visual and Interaction Settings: Customizable settings allow users to adjust the tool according to their specific needs. This could include adjustable font sizes, high-contrast color options for better visibility, and simplified user interfaces for users with cognitive impairments. For instance, a user with color blindness could enable a high-contrast mode or grayscale option, while someone with a cognitive disability might prefer simplified text and fewer complex visualizations. These adjustments are essential to making the tool accessible to as many users as possible.
- Speech Recognition and Assistive Technologies: The tool should also integrate with assistive technologies such as speech recognition or dictation software, allowing users to interact with the tool using voice commands. For those with limited motor function, integrating compatibility with alternative input devices such as switches or head pointers can enhance usability. Such inclusive features ensure that people with a wide range of abilities can fully engage with the tool, breaking down barriers to access.
- Low-Tech and Low-Bandwidth Support: Inclusivity must also extend to users in areas with limited access to high-end technology or stable internet connections. The tool should be optimized to function on older devices and with low-bandwidth connections, ensuring it can still provide accurate and timely insights even in regions with technological constraints. This allows for broader global adoption, particularly in developing countries or remote regions where access to cutting-edge technology is limited.

3.3 CULTURAL SENSITIVITY

- Avoiding Cultural Assumptions: Developers must avoid making assumptions about users' cultural or socio-economic backgrounds. The design of the tool should not reflect a bias toward any particular cultural norms or assumptions about literacy, access to technology, or political and social attitudes. For example, not all users will have a high level of literacy or familiarity with digital tools, so the interface should be designed to be intuitive and simple enough for users of all backgrounds.
- Culturally Relevant Content Detection: The tool's detection algorithms should account for culturally specific disinformation. For instance, in some regions, disinformation might exploit historical or religious tensions, while in others, it might focus on ethnic or racial divides. AI models must be trained to recognize these regional and cultural nuances, ensuring the tool is equally effective in detecting content that may only be relevant to specific locales. For example, disinformation about vaccines might take different forms in North America versus Africa versus Europe, where cultural beliefs and healthcare access differ significantly.

Funded by the European Union

- Inclusive User Testing: During the development process, the tool should undergo extensive testing with diverse user groups to ensure that it is truly inclusive. Beta testing should involve participants from a variety of linguistic, cultural, and socio-economic backgrounds. This will help developers identify and address any potential barriers to access or bias that may arise from the tool's design or functionality. Input from these diverse user groups should be continuously incorporated to enhance the tool's adaptability and inclusiveness
- Cultural Neutrality in User Interface: The icons, symbols, images, and even color choices used in the interface must be culturally neutral and easily understood by users from different regions. For example, certain colors or symbols that are positive in one culture may have negative connotations in another. The design must avoid reinforcing any cultural stereotypes or assumptions, and imagery used in the tool should be appropriate for a global audience. This ensures that the tool resonates with users from various cultural contexts without causing offense or confusion.

3.4 IMPLEMENTATION RECOMMENDATIONS

1. Regional Partnerships for Data Collection: Collaborate with local experts or institutions in different regions to collect culturally and linguistically relevant data that reflects local disinformation trends. These partnerships will help train AI models that are context-sensitive and regionally accurate.
2. Standards for Accessibility: Ensure that the tool adheres to global accessibility standards like the Web Content Accessibility Guidelines (WCAG). These standards provide a framework for creating digital content that is accessible to people with disabilities, covering everything from color contrast ratios to screen reader compatibility
3. Feedback Loops from Diverse Users: Implement continuous feedback loops where users from different cultural and linguistic backgrounds, as well as individuals with disabilities, can provide insights into how the tool functions in their specific contexts. This feedback should guide future iterations and updates of the tool, ensuring ongoing improvement in inclusivity.

3.5 CONCLUSION

Inclusivity in AI-driven disinformation tools is critical to ensuring their effectiveness and adoption across different global contexts. By supporting multiple languages, making the tool accessible to users with disabilities, and designing it to be culturally sensitive, developers can create a platform that works for a diverse range of users. From integrating assistive technologies to considering regional disinformation patterns, the tool should be adaptable to meet the unique needs of all users, regardless of their technical expertise, physical abilities, or cultural background.

Funded by the European Union

4 FOCUS ON SOCIAL MEDIA

Social media platforms are a primary battleground for the spread of disinformation and misinformation, largely due to their speed, reach, and engagement mechanisms. For AI-driven disinformation detection tools to be effective, they must focus on the specific features, trends, and behaviours that define how content spreads across social media. This section explores how developers can ensure that their tools are well-equipped to analyze, detect, and report disinformation in real-time by fully integrating with social media platforms and accounting for the unique ways in which content proliferates online.

4.1 INTEGRATION WITH SOCIAL MEDIA APIS

- Real-Time Analysis through API Integration: One of the most critical features of an AI disinformation tool is its ability to access and analyze data in real-time. To achieve this, the tool must integrate seamlessly with the APIs (Application Programming Interfaces) of major social media platforms, such as Facebook, Twitter, Instagram, TikTok, and YouTube. These APIs provide access to a wide array of data, including public posts, user interactions, engagement metrics, and content-sharing patterns. Through API integration, the tool can constantly monitor social media content, identify emerging disinformation trends, and flag suspicious posts as they go viral. API integration can however be a challenging issue. They often provide limited access to the platform data. Also they can change the APIs codes without any further warning, making the codes and connection obsolete for the programing.
- Cross-Platform Compatibility: The tool should be designed to work across multiple platforms, given that disinformation rarely remains isolated to just one. For example, a piece of disinformation may first appear on Twitter, then quickly spread to Facebook, Instagram, and even more niche platforms like Telegram. By integrating with the APIs of all these platforms, the tool can track the lifecycle of a disinformation campaign across different media channels, providing a more holistic view of how content spreads.
- Deep Access to Metadata: When integrating with social media APIs, the tool should prioritize accessing and analyzing metadata, such as timestamps, geolocation data (where available), the original source of posts, and user engagement. This data is crucial for understanding the virality and reach of posts, as well as for identifying whether a post is being artificially amplified by bots or coordinated inauthentic behaviour. By examining metadata, the tool can build a more comprehensive picture of how and why certain content gains traction.
- Privacy Considerations in API Use: While social media APIs provide access to a wealth of data, the tool must ensure compliance with privacy laws such as the General Data Protection Regulation (GDPR) in the EU and other regional privacy standards. This means anonymizing user data where possible, limiting the scope of data collection to public information, and providing users with clear insights into how their data is being used by the tool. Ensuring that the tool respects privacy standards will help prevent misuse and build trust with users and stakeholders.

Funded by the European Union

4.2 STAY UPDATED ON TRENDS

- Tracking Platform-Specific Features: Each social media platform operates differently in terms of how content is shared, what types of posts receive the most engagement, and what algorithms are used to promote or demote content. For example, Twitter emphasizes trending hashtags and retweets, while Instagram focuses on visual content like images and stories, and TikTok amplifies short-form videos using personalized "For You" feeds. The tool must be designed to track these platform-specific features and analyze content accordingly.
    - Example 1: On Twitter, the tool should monitor trending hashtags, identifying those that are linked to coordinated disinformation efforts.
    - Example 2: On Instagram, it could analyze image-based disinformation, which might involve altered or misleading images accompanied by manipulated captions.
    - Example 3: On TikTok, the tool should detect viral videos that use deceptive editing or misleading claims to promote disinformation narratives.

By keeping up with the constantly evolving algorithms and features of each platform, the tool can ensure it remains relevant and accurate in analyzing new types of content.

- Adapting to Algorithm Changes: Social media platforms frequently update their algorithms to promote or suppress certain types of content, often in response to public pressure or policy changes. These algorithmic changes can significantly impact how disinformation spreads. The tool must be continuously updated to adapt to these changes, ensuring that it can still accurately detect content even when platforms alter their content distribution methods. For example, a platform might prioritize content from verified sources, reducing the visibility of disinformation, or it might introduce new features like stories or fleets, which require the tool to analyze a different type of content format.
- Monitoring Emerging Platforms: In addition to mainstream platforms like Facebook, Twitter, and Instagram, the tool should be capable of integrating with and analyzing emerging social media platforms. Disinformation campaigns often shift to newer platforms when traditional ones tighten their content moderation policies. For example, platforms like Gab, Parler, and Telegram have seen increased disinformation activity due to their more lenient moderation policies. By staying updated on these newer platforms, the tool can track disinformation migration from one platform to another and provide a more comprehensive analysis.

4.3 VIRAL SPREAD AND ENGAGEMENT

- Understanding Viral Mechanics: One of the defining features of social media is the potential for content to go viral, reaching millions of users in a short amount of time. Disinformation often takes advantage of these viral mechanisms to spread quickly and widely before it can be effectively

Funded by the European Union

countered. The tool must be equipped to analyze how content goes viral, taking into account the engagement metrics unique to each platform, such as likes, shares, retweets, comments, and replies.

  - Engagement Metrics as Indicators: High engagement does not always indicate trustworthiness; in fact, disinformation often garners significant engagement because it is sensational or emotionally charged. The tool should be designed to prioritize high-engagement content for analysis, as this content is more likely to have gone viral and thus has the greatest potential for widespread disinformation. By analyzing posts that have received an unusually high number of likes, shares, or retweets in a short period, the tool can quickly identify potential disinformation campaigns.

- Network Analysis of Spread: The tool should not just track engagement numbers, but also map out the network of interactions that help disinformation spread. For example, it could identify key influencers (whether human or bot) who are responsible for amplifying false narratives. This type of network analysis can reveal the hubs of disinformation and provide insights into how fake news propagates across social media.

  - Example: The tool might detect that a small number of accounts are responsible for a disproportionate amount of retweets or shares, indicating that these accounts may be part of a coordinated disinformation campaign.

- Analyzing Content Lifecycle: In addition to viral spread, the tool should analyze the lifecycle of disinformation on social media. This includes tracking when a post first appears, how quickly it spreads, and when it reaches its peak engagement. The tool could provide insights into whether certain types of disinformation (e.g., political vs. health-related) have longer lifespans on social media than others. It could also identify patterns in how disinformation campaigns resurface, perhaps during politically or socially sensitive periods, such as elections or public health crises.

4.4 IMPLEMENTATION RECOMMENDATIONS

1. Deep API Integration: Work with social media platforms to ensure the tool has deep integration with their APIs, allowing access to real-time data on posts, engagement, and user interactions. It would be interesting here to ask for API stability and a minimum set of accessible data at the EU level.

2. Constantly Update for Platform Changes: Allocate resources for a team that monitors platform updates and trends in disinformation, ensuring the tool adapts to new content types, engagement metrics, and algorithms introduced by platforms. However, the priority is for the teams to ensure that the tool is working well on a continuous basis whatever programs are used, and the platform strategy should be only an available option, to enhance some processing.

3. Leverage Advanced Network Analytics: Integrate network analysis tools that can map the spread of disinformation and identify key influencers or bots driving the amplification of fake news.

4.5 CONCLUSION

The focus on social media is crucial for any AI disinformation detection tool. By integrating with social media APIs, keeping up with platform-specific trends, and understanding the mechanics of viral content, the tool can stay effective in detecting and flagging disinformation in real-time. Social media is constantly evolving, and disinformation campaigns adapt quickly to take advantage of new features and loopholes. Therefore, tools that focus on social media must be nimble, adaptive, and always one step ahead in analyzing the flow and impact of disinformation across various platforms.

Funded by the European Union

5 TACKLING COORDINATED INAUTHENTIC BEHAVIOUR

Coordinated inauthentic behaviour (CIB) refers to the deliberate use of fake or compromised accounts, bots, or organized groups to manipulate social media platforms by amplifying disinformation, creating artificial trends, or influencing public opinion. Tackling CIB is critical for AI-driven disinformation tools, as these behaviours can rapidly spread false information and manipulate online discourse. To effectively detect and neutralize CIB, developers must employ advanced techniques such as network analysis, pattern recognition, and behaviour flagging. CIB can be monitored by tool developers using a combination of techniques that allow for easy identification, tracking, and mitigation of suspicious activities on social media platforms. The goal is to empower users to understand and potentially act upon insights provided by the AI tools. Below is a detailed breakdown of how developers can monitor and address CIB from a user perspective, ensuring that users are provided with transparent and actionable information.

5.1 PROVIDING REAL-TIME ALERTS ON SUSPICIOUS ACTIVITY

- Notification System: For the user, a key feature is real-time alerts when CIB is detected. The tool should monitor social media platforms continuously and notify users when unusual or suspicious behavior is observed, such as a sudden spike in engagement on a particular post or the coordinated use of hashtags. This notification system should be customizable, allowing users to set thresholds for alerts based on engagement metrics or content types they are interested in monitoring.
    - Example: A user tracking political disinformation could receive an alert when a newly posted article about an election is retweeted hundreds of times within minutes by accounts that seem unrelated.
- Dashboards for Monitoring: Developers should provide users with interactive dashboards that display real-time metrics related to CIB, such as engagement spikes, bot activity, and coordinated hashtag campaigns. This dashboard should be user-friendly and customizable, allowing users to filter and focus on specific platforms, hashtags, or accounts they want to monitor for suspicious activity. The data should be presented in easy-to-understand visualizations, like graphs or charts, to help users quickly identify anomalies.
    - Example: Users could view a graph of a hashtag's activity over time, with spikes clearly marked as potentially coordinated behavior if they happen unnaturally fast or involve certain flagged accounts.

Funded by the European Union

5.2 EDUCATING USERS ON IDENTIFIABLE CIB PATTERNS

- Pattern Recognition Training: From a user perspective, understanding the patterns of CIB is crucial. The tool should offer educational resources that explain how coordinated accounts typically behave, such as posting at the same time, using identical or very similar language, or promoting the same narrative across multiple platforms simultaneously. By educating users on what to look for, the tool helps them independently verify suspicious activity without needing advanced technical knowledge.
    - Example: A guide within the tool could explain that accounts using a specific hashtag in rapid succession, often at identical intervals, could be a sign of coordination, especially if these accounts have limited posting history or were created recently.
- Flagging Behavior for Users: The tool should automatically flag behaviors that are typical of CIB. For instance, if a group of accounts consistently retweets or shares posts within seconds of each other, the tool could flag this as potentially coordinated behavior and present it to the user in a straightforward way, such as highlighting these posts in a separate section of the dashboard.
    - Example: Users monitoring specific keywords could be notified when multiple accounts use the same phrase in a coordinated way across platforms, indicating the potential for a bot-driven campaign.

5.3 OFFERING NETWORK VISUALIZATION TOOLS

- Visualizing Networks: Developers should provide tools that allow users to visualize the networks of accounts involved in spreading disinformation. For instance, a user should be able to see a network map of accounts interacting with specific posts or hashtags, where connections between accounts are shown based on their interaction frequency. This makes it easier for users to spot clusters of accounts that might be working together inauthentically.
    - Example: A user could view a network graph showing that a small group of accounts retweeted the same disinformation post at an abnormally high frequency, with lines connecting these accounts to show their interactions.
- Detecting Influence Patterns: The tool should also highlight the key influencers within these networks—accounts that are central to disinformation campaigns. By identifying the accounts that are most effective at amplifying disinformation, the tool can help users understand the source of the coordinated activity and whether it is being driven by bots, fake accounts, or influential individuals.
    - Example: In a visual network of retweets, users can easily see if a few central accounts (influencers) are responsible for the majority of interactions, suggesting coordinated amplification by a few key players.

Funded by the European Union

5.4 ENABLING USER-DRIVEN INVESTIGATIONS

- Custom Search and Monitoring Features: Users should be able to input specific queries into the tool to monitor potential CIB activities around particular topics, keywords, or hashtags. This level of control empowers users to investigate specific disinformation narratives that may concern them and observe any coordinated efforts to amplify those narratives.
    - Example: A journalist could set up the tool to track discussions around a particular political event, observing whether a sudden flood of posts using similar hashtags or phrases occurs around the same time, suggesting coordinated efforts to spread a narrative.
- Historical Analysis: Users should also be able to perform retrospective analyses of disinformation campaigns, tracking how certain content spread over time and whether it was promoted by coordinated efforts. The tool could offer a feature that lets users see how the disinformation started, how quickly it gained traction, and which accounts were instrumental in spreading it.
    - Example: A researcher could use the tool to analyze a disinformation campaign from the previous year, identifying the timeline of posts, the key accounts involved, and the moments when coordinated behavior spiked.

5.5 DETECTING BOTS AND FAKE ACCOUNTS FOR USERS

- Bot and Fake Account Identification: The tool should automatically detect and label bots or fake accounts, using predefined criteria such as abnormal posting frequency, newly created accounts, generic profile pictures, or a lack of personal interaction on the account. These accounts should be flagged to the user, who can then decide whether to disregard content from these accounts or investigate them further.
    - Example: When a user comes across content that has been shared by numerous accounts, the tool can highlight how many of these are bots or fake accounts, allowing the user to easily see how much of the engagement is artificial.
- User-Driven Flagging: Users should also have the ability to flag accounts or behaviors they believe to be suspicious. This could include accounts that post content repetitively, interact in highly formulaic ways, or suddenly appear in discussions about trending topics. The tool should enable users to flag such behavior and provide them with feedback on whether the flagged account matches known patterns of CIB.
    - Example: If a user notices an account repeatedly posting similar content with multiple identical hashtags in short bursts, they could flag it. The tool would then provide an analysis of whether this behavior aligns with typical bot activity.

Funded by the European Union

5.6 CROSS-PLATFORM MONITORING AND COORDINATION

- Multi-Platform Tracking: CIB activities often span multiple social media platforms, as disinformation campaigns seek to maximize reach and engagement. Developers should ensure that the tool can track disinformation across platforms, showing users how coordinated activity on one platform (e.g., Twitter) is mirrored or amplified on another (e.g., Facebook). This gives users a complete picture of how disinformation spreads beyond the boundaries of a single platform.
    - Example: A user could track how a hashtag campaign that begins on Twitter gains traction on Instagram or how a YouTube video promoting disinformation is shared across multiple platforms at once.
- Timeline of Coordinated Campaigns: The tool should allow users to view the timeline of a coordinated campaign, showing when and how specific pieces of disinformation appeared on different platforms. This timeline feature would help users identify key moments when a campaign gained momentum and assess whether it was driven by inauthentic coordination across platforms.
    - Example: A user tracking climate change misinformation could see how the same content appeared on multiple platforms within minutes or hours, indicating a highly coordinated effort to spread the false narrative.

5.7 TRANSPARENCY AND FEEDBACK

- Explaining Results to Users: For users to trust the tool's analysis of CIB, they need to understand how the tool reaches its conclusions. Developers should include detailed explanations of why certain accounts, posts, or behaviors are flagged as inauthentic or coordinated. This transparency will help users feel confident in the tool's findings and empower them to act on the information provided.
    - Example: If a tool flags a group of accounts as being part of a bot network, it should explain which behaviors (e.g., posting frequency, identical interactions) led to this conclusion, allowing the user to verify the tool's analysis.
- User Feedback Mechanisms: The tool should provide a way for users to give feedback on the flagged content or behavior. If users believe that certain behavior has been incorrectly flagged or that something suspicious was missed, they should be able to provide feedback that the tool's algorithms can learn from. This feedback loop will improve the tool's ability to detect CIB over time.
    - Example: A user could submit feedback if they believe a flagged account is actually legitimate or if they notice suspicious behavior that the tool did not detect. The tool could then incorporate this input into future detection efforts.

Funded by the European Union

5.8 RECOMMENDATIONS FOR FACT CHECKERS AND MEDIA PROFESSIONALS

1. Real-Time Alerts: Develop a customizable notification system that alerts users to suspicious CIB activities, such as spikes in engagement or coordinated hashtag use, with flexible thresholds based on user preferences.
2. Interactive Dashboard: Provide a user-friendly dashboard displaying real-time metrics like bot activity and coordinated campaigns. Include customizable graphs and network maps to help users spot anomalies quickly.
3. User Education: Offer educational resources and guides explaining common CIB patterns (e.g., simultaneous posting, repetitive language), enabling users to independently verify suspicious behavior.
4. Network Visualization: Implement network mapping tools to visualize connections between accounts involved in spreading disinformation, highlighting key influencers driving CIB efforts.
5. Custom Search and Monitoring: Allow users to input specific queries for monitoring CIB around keywords or hashtags, and offer retrospective analysis of past disinformation campaigns.
6. Bot Detection: Automatically detect and label bot and fake accounts, and provide users with the option to flag suspicious accounts for review.
7. Cross-Platform Tracking: Enable tracking of disinformation across multiple platforms, with a timeline feature showing how campaigns gain traction across different social media sites.
8. Transparent Flagging: Provide detailed explanations for flagged CIB behavior, and allow users to contest flagged content to improve the tool's accuracy.
9. User Feedback: Incorporate user feedback into regular updates to refine the tool, with mechanisms for users to report false positives and suspicious content.

5.9 CONCLUSION

From a user perspective, monitoring CIB should be an intuitive and transparent process facilitated by the AI tool. Through features such as real-time alerts, network visualization, custom searches, and bot detection, users can stay informed about suspicious activities on social media. Educating users on common CIB patterns and giving them the ability to flag suspicious behavior allows for an interactive and user-centric experience. By developing these features, tool developers can empower users to better understand and mitigate the impact of CIB on public discourse.

Funded by the European Union

6 ETHICS

Ethical considerations are paramount when developing AI tools designed to monitor, detect, and mitigate disinformation. These tools must balance the need for effective disinformation detection with a commitment to upholding user privacy, transparency, and fairness. Additionally, developers must ensure that the tools are not misused to suppress legitimate voices or content. Ethical AI ensures that the technology works equitably and transparently while respecting the rights and interests of users. Below is an in-depth exploration of the key ethical principles that should guide developers in creating tools to tackle disinformation and coordinated inauthentic behavior.

6.1 DATA PRIVACY AND SECURITY

- User Privacy First: Given that AI tools for disinformation detection often involve the collection and analysis of vast amounts of social media data, protecting user privacy is essential. Developers must ensure that they adhere to global data protection standards, such as the General Data Protection Regulation (GDPR) in Europe or other applicable regional privacy laws. This involves clearly defining what data the tool collects, how it is processed, and ensuring that data collection practices are proportional to the goals of disinformation detection.
    - Avoid Over-Collection of Data: The tool should only collect data that is strictly necessary for its functioning. For example, if the tool is designed to analyze disinformation in public posts, it should not gather unnecessary private data, such as private messages, user browsing habits, or personal information unless explicitly required by the task and consented to by the user. Over-collection increases the risk of data misuse and breaches, which could harm users and damage the tool's credibility.
    - Anonymization of Data: Developers must ensure that any personal data that is collected is anonymized to the greatest extent possible, particularly in cases where sensitive data such as location or political opinions are being monitored. This protects users from being personally identified based on their activity within the tool, further safeguarding their privacy.
    - Security of Data Handling: Strong cybersecurity measures must be in place to ensure the safe storage and transmission of any data collected by the tool. This includes employing encryption techniques, secure access controls, and regular audits to prevent unauthorized access or data breaches. Developers should also establish clear protocols for data handling, ensuring that only authorized personnel can access sensitive information.
    - Clear and Accessible Privacy Policies: Developers must provide clear and comprehensive privacy policies that explain what data is collected, why it is collected, how it will be used, and for how long it will be retained. This policy should be easily accessible to users before

Funded by the European Union

they interact with the tool. Additionally, users should have the ability to opt out of data collection where feasible, particularly for non-essential features.

6.2 TRANSPARENCY

- Transparent Methodologies: Transparency in how the tool operates is critical for building user trust. Developers must ensure that the tool's algorithms, methodologies, and decision-making processes are transparent and clearly communicated to users. This involves providing documentation or explainers that detail how disinformation is identified, how scores or flags are assigned, and the criteria used to determine whether content is classified as fake or manipulated.
    - Explanation of Detection Methods: For example, if the tool uses machine learning to detect disinformation, users should be able to access an overview of how the model works, including the types of data it uses (e.g., engagement metrics, language patterns, network behavior), and any limitations of the system. This ensures users understand that while the tool is powerful, it may still produce false positives or false negatives.
- Option to Contest Results: Users must have the ability to contest the results or decisions made by the tool. If content is flagged as disinformation or accounts are suspected of engaging in coordinated inauthentic behavior, users should be provided with an opportunity to appeal or challenge these findings. This could be done by offering a simple feedback or dispute mechanism within the tool, where users can submit evidence that the flagged content or activity is legitimate. This appeals process not only promotes fairness but also helps improve the tool by correcting potential errors.
    - Example: A legitimate activist account might be wrongly flagged as part of a coordinated campaign because it uses popular hashtags to promote social justice causes. In this case, the user should be able to contest the flagging by explaining the nature of their activism and providing supporting evidence.
- Limitations of the Tool: It is essential that the tool is clear about its limitations. While AI is a powerful tool for detecting patterns and anomalies, it is not infallible. Developers should communicate the tool's boundaries, explaining that it may not detect all disinformation or CIB efforts and that it is still subject to refinement and updates. This ensures that users are aware of the tool's scope and don't develop false confidence in its ability to detect all malicious activities.

6.3 AVOIDING BIAS

- Regular Bias Audits: One of the major ethical challenges in AI development is the potential for bias in content analysis. This can occur if the tool is trained on biased data that overrepresents certain groups or viewpoints, leading to the disproportionate flagging of content from specific regions,

Funded by the European Union

communities, or demographics. To avoid this, developers must conduct regular bias audits to identify and correct any algorithmic bias that might unfairly target particular groups.

  - Diverse Data Sources: AI models should be trained on diverse datasets that represent a wide range of cultural, geographic, and linguistic perspectives. This reduces the likelihood of the tool disproportionately flagging content from certain regions or social groups, ensuring that the tool operates fairly across a global user base.
  - Addressing Socio-Political Biases: Developers should also be vigilant in preventing the tool from reflecting political or ideological biases. For example, if the tool disproportionately flags content from activist groups or political dissidents, it could be seen as suppressing legitimate expression. Developers need to ensure that the tool does not become a tool of censorship or political manipulation.

- Real-Time Monitoring and Adjustments: As the tool is used in real-world scenarios, developers should implement real-time monitoring of its outputs to detect whether certain groups or viewpoints are being unfairly targeted. If biases are detected, adjustments to the algorithms should be made immediately to ensure a more balanced approach to disinformation detection.
  - Example: If the tool flags a higher proportion of posts from non-Western regions as disinformation, developers should investigate whether the training data includes enough examples from those regions or whether cultural differences are being incorrectly interpreted as signs of disinformation.

6.4 ETHICAL SAFEGUARDS

- Preventing Misuse of the Tool: While the tool is designed to detect disinformation and CIB, it could be misused to suppress legitimate voices or target certain individuals or groups. Developers must implement ethical safeguards to prevent this from happening. For example, if governments or organizations gain access to the tool, they should not be able to use it to silence dissent or opposition under the guise of disinformation detection.
  - Content Moderation Ethics: When flagging content for moderation, the tool should incorporate ethical standards that ensure it does not unduly suppress legitimate free speech. Developers should collaborate with human moderators to ensure that flagged content is reviewed within an ethical framework, balancing the need to remove harmful disinformation while protecting free expression.
  - Safeguarding Against Government Abuse: If the tool is made available to government entities, strict terms of use should be in place to prevent misuse. This includes prohibiting the tool's use for political surveillance, targeting political opponents, or censoring content that falls outside the scope of harmful disinformation. Developers must ensure that the tool's power to flag disinformation is not exploited to oppress vulnerable communities or silence dissent.

Funded by the European Union

- Mitigating the Impact of False Positives: Ethical AI tools should be designed with mechanisms that mitigate the potential damage caused by false positives, where legitimate content is incorrectly flagged as disinformation. Developers can address this by ensuring that users are notified if their content is flagged and are provided with an explanation and appeal process. In cases where legitimate content is flagged, efforts should be made to rectify the situation quickly and transparently.
    - Example: If a journalist's report is wrongly flagged as disinformation because it contains sensitive political content, the tool should allow for quick and easy rectification once the error is brought to light.

6.5 RECOMMENDATIONS FOR MEDIA PROFESSIONALS AND TOOLS' DEVELOPERS

1. Clear Ethical Guidelines and Privacy Policies: Develop clear, accessible privacy policies and ethical guidelines that explain data handling, usage rights, and the responsibilities of both developers and users.
2. Ongoing Bias and Ethics Audits: Implement regular audits to monitor for algorithmic bias and assess the tool's ethical impact on marginalized or vulnerable groups. These audits should be transparent and shared with users to build trust in the tool's fairness.
3. Collaboration with Ethics Committees: Establish an independent ethics committee to oversee the development and deployment of the tool. This committee can help evaluate the tool's adherence to ethical standards, particularly in areas such as data privacy, bias, and human rights.
4. User Empowerment and Control: Provide users with control over their data and the ability to contest flagged content. Users should be given the tools to understand why their content or activity was flagged and a clear process for disputing those decisions.

6.6 CONCLUSION

Developers must prioritize ethics when designing AI tools for disinformation detection. This involves upholding data privacy, ensuring transparency, avoiding bias, and implementing safeguards against the misuse of the tool. By maintaining a strong ethical framework, developers can ensure that the tool protects users' rights, promotes fairness, and builds trust among stakeholders. Ethical AI is not just about building effective tools—it's about ensuring that these tools operate with integrity and respect for human dignity.

Funded by the European Union

7 INTEGRATING LANGUAGES

As disinformation campaigns are often global in nature, AI tools designed to detect and counter disinformation must be equipped to handle multilingual environments. Language plays a crucial role in how information is shared and how disinformation spreads across different regions and cultural contexts. Integrating robust multilingual support into disinformation detection tools is essential for ensuring the tool's effectiveness across different countries, regions, and language groups. Additionally, the ability to recognize cultural sensitivities and adapt to local nuances makes the tool more relevant and impactful in diverse settings. Below is a detailed elaboration on how integrating languages can improve the functionality and reach of AI tools for disinformation detection.

7.1 MULTILINGUAL SUPPORT

- Language-Specific Models for Greater Accuracy: To ensure that the tool is effective across multiple languages, it must use language-specific models that account for the unique syntax, grammar, and colloquial expressions found in each language. Disinformation in different languages can manifest in ways that are distinct from each other, and direct translation may not capture the nuances of a particular language. For example, idiomatic expressions, slang, or culturally specific references in one language may not have direct equivalents in another. Using models that are tailored to the language being analyzed allows the tool to more accurately detect disinformation by recognizing patterns unique to that language.
    - Example: In Arabic, disinformation might be spread using certain phrases that leverage religious or historical references. A language-specific model for Arabic could be trained to recognize these patterns, whereas a generic or translated model might miss them.
- Natural Language Processing (NLP) Across Languages: The tool must employ natural language processing (NLP) models that are capable of analyzing content in different languages without losing the accuracy and depth of the analysis. This involves using NLP techniques such as tokenization, named entity recognition (NER), and sentiment analysis that are adapted to each specific language. The accuracy of these tools depends on the quality and diversity of the language data used to train the models, so it is critical to ensure that the datasets include a wide variety of language styles, dialects, and regional variations.
    - Example: In Spanish, the way disinformation spreads in Spain may differ from how it spreads in Latin America, due to differences in regional vocabulary and political contexts. By using regionalized language models, the tool can detect disinformation in both regions with greater accuracy.
- Machine Translation with Cultural Context: In some cases, it might be necessary to analyze content in languages for which there isn't sufficient NLP model support. In such cases, machine translation can be used, but developers should ensure that these translations are contextual and culturally

Funded by the European Union

accurate. Machine translation models should not only convert text from one language to another but also take into account cultural nuances to avoid misinterpretation of meaning. This is especially important for sensitive topics where mistranslations could alter the intent of the content.

- Example: Translating a politically charged statement from Mandarin to English requires attention to both the linguistic structure and the political context in which the statement was made to ensure the message's nuances are not lost or misunderstood.

7.2 CROSS-LANGUAGE ANALYSIS

- Comparing Disinformation Across Languages: One of the most valuable features in a multilingual disinformation detection tool is the ability to perform cross-language analysis. This allows users to compare insights and trends across different languages, offering a broader perspective on how disinformation spreads globally and how narratives evolve as they move from one language to another. Cross-language analysis can help identify patterns of disinformation that may begin in one region or language and then spread to others with slight variations.
  - Example: A piece of disinformation about a global health issue (such as vaccine misinformation) might originate in English and later appear in French, Spanish, and other languages. By comparing how the misinformation is translated, localized, and amplified in each language, the tool can provide a deeper understanding of how disinformation strategies adapt across cultures.
- Tracking Narrative Shifts Across Regions: Disinformation narratives often evolve as they are translated into different languages and adapted to local contexts. A robust tool should allow users to track these narrative shifts across regions, identifying how certain themes or messages are altered to appeal to local audiences. This feature is particularly useful for global organizations, researchers, and policymakers who need to understand the global reach and impact of disinformation.
  - Example: A political disinformation campaign that starts in Russian media might be translated into English and adjusted to reflect concerns specific to English-speaking audiences, such as focusing on particular political candidates or issues. The tool could track these shifts in messaging and analyze how the core disinformation theme changes as it crosses language barriers.
- Cross-Language Correlation of Hashtags and Keywords: Disinformation campaigns often use hashtags or keywords that are translated or adapted into different languages to maintain the same theme across regions. The tool should be able to correlate these keywords and hashtags across languages to identify multi-lingual disinformation campaigns. This feature would help users see whether certain hashtags, phrases, or keywords are part of a larger coordinated campaign that spans multiple languages.

Funded by the European Union

- Example: A disinformation campaign about climate change might use specific hashtags in English, but those hashtags might have counterparts in other languages (e.g., Spanish, French). The tool could analyze how the same campaign spreads using equivalent hashtags in different languages.

7.3 CULTURAL SENSITIVITY

- Recognizing Cultural Nuances in Disinformation: Disinformation often relies on culturally specific references, narratives, and biases to gain traction within certain communities. Therefore, it is essential that the tool is not just multilingual but also culturally sensitive, meaning it can recognize and understand the nuances of how disinformation manifests in different regions and cultural contexts. Cultural differences in how people perceive news, politics, and social issues can significantly impact how disinformation is framed and spread. The tool must be trained to detect these subtleties in order to accurately flag and analyze disinformation in different regions.
    - Example: Disinformation in the Middle East might exploit religious narratives or historical tensions, while disinformation in Western Europe might focus more on political or economic concerns. The tool should be equipped to detect these different narratives and understand their cultural significance.
- Localization of Disinformation Patterns: Developers should ensure that the tool is capable of localizing disinformation detection by training it on region-specific datasets that reflect local dialects, cultural references, and socio-political dynamics. Localization ensures that the tool is not simply applying a one-size-fits-all approach but is instead tailored to the unique characteristics of each region. For instance, a disinformation narrative about government corruption might look different in South America than it does in Europe, and the tool must be able to adapt to these differences.
    - Example: In Brazil, disinformation campaigns might leverage cultural references related to football or Carnival to make false claims more relatable to local audiences, while in the U.S., campaigns might focus on issues like gun control or healthcare. The tool needs to be sensitive to these cultural factors when analyzing content.
- Avoiding Cultural Bias in Analysis: While recognizing cultural nuances, the tool must also avoid cultural bias in its analysis. Cultural biases can arise if the tool disproportionately flags content from certain regions or cultures as disinformation due to misunderstandings of local norms, behaviors, or language usage. To counter this, the tool should be trained on diverse datasets that include a wide range of cultural contexts, ensuring that it does not unfairly target specific communities or misinterpret culturally specific content as disinformation.
    - Example: A tool trained primarily on Western datasets might incorrectly flag content from indigenous communities that use specific language or cultural references unfamiliar to

Funded by the European Union

mainstream Western discourse. Ensuring diverse training datasets helps avoid such misinterpretations.

- Handling Linguistic Ambiguities: Languages often contain ambiguities and double meanings that can change the interpretation of a statement. Developers must ensure that the tool is able to handle these ambiguities by taking into account the context in which words are used. This is particularly important in languages with homophones or languages where tone and intonation significantly affect meaning (e.g., Chinese, Thai). By integrating contextual analysis into the tool's language models, developers can reduce the risk of false positives and improve accuracy in disinformation detection.

7.4 IMPLEMENTATION RECOMMENDATIONS

1. Develop Language-Specific Models: Collaborate with linguists, native speakers, and local experts to develop language-specific models that account for the linguistic structure and cultural nuances of each language the tool supports.
2. Enhance Machine Translation for Context: For languages where NLP models are not yet fully developed, use context-sensitive machine translation systems that preserve the cultural and linguistic integrity of the original message.
3. Cross-Language Keyword and Hashtag Correlation: Build a system that tracks cross-language correlations between keywords, hashtags, and narratives, allowing users to see how disinformation spreads across regions and languages.
4. Train Models with Culturally Diverse Data: Ensure that the AI models are trained on culturally diverse datasets to avoid cultural biases in the detection process and to better understand local contexts of disinformation.

7.5 CONCLUSION

Integrating languages into AI-driven disinformation detection tools is more than just adding translation capabilities—it involves building a system that is linguistically and culturally aware. By enabling multilingual support, providing cross-language analysis, and incorporating cultural sensitivity, developers can create a tool that is effective across diverse regions and contexts. The ability to accurately detect disinformation in multiple languages, track its spread across different linguistic groups, and understand the cultural nuances behind disinformation narratives will greatly enhance the tool's global impact. By ensuring the tool remains sensitive to local contexts while maintaining linguistic accuracy, developers can help users combat disinformation in a more comprehensive and targeted manner.

Funded by the European Union

8 EXPLAINING "FAKENESS"

One of the most crucial elements of AI tools designed to detect and flag disinformation is the ability to explain the "fakeness" of content in a clear, transparent, and actionable manner. For users to trust the system and understand its decisions, the tool must provide detailed insights into how and why certain content is classified as fake or manipulated. This involves a multi-factor scoring system that evaluates the likelihood of disinformation, clear explanations for flagged content, and resources that empower users to identify disinformation independently. Below is an in-depth elaboration on how developers can integrate these elements to improve user understanding of the "fakeness" of content.

8.1 DISTINGUISHING BETWEEN DISINFORMATION AND MISINFORMATION

- Disinformation: Refers to false or manipulated information that is intentionally spread to deceive or mislead. Disinformation is often used as part of coordinated campaigns, typically to achieve political, financial, or social goals. It involves deliberate efforts to shape perceptions, influence public opinion, or undermine trust in institutions by spreading content that is known to be false. Disinformation is often amplified by bots, fake accounts, or coordinated inauthentic behavior to create the illusion of widespread support or consensus around a false narrative.
    - Example: A group of fake social media accounts spreads a fabricated story about a political candidate engaging in illegal activities. The intent is to damage the candidate's reputation ahead of an election, and the accounts actively coordinate to amplify this false story to reach as many people as possible.
- Misinformation: Refers to false or misleading information that is spread unintentionally. In contrast to disinformation, misinformation is shared by individuals or groups who believe the content to be true but are misinformed. It can often arise from misunderstandings, incorrect interpretations, or poorly sourced information. While the content itself may be false, there is no malicious intent behind its dissemination. Nevertheless, misinformation can still cause significant harm, particularly when it spreads widely and influences public opinion.
    - Example: A social media user shares an outdated or misinterpreted article about a health issue, believing it to be accurate. While the user has no intent to deceive others, the spread of this misinformation could lead to confusion or harm.

8.2 DETECTING AND SCORING DISINFORMATION VS. MISINFORMATION

- Intent Detection: One of the challenges for AI tools is determining the intent behind the spread of false information. While intent can be difficult to assess directly, the tool can look for patterns of behavior that indicate whether the content is being deliberately spread as part of a

Funded by the European Union

coordinated campaign (disinformation) or whether it appears to be shared organically by individuals (misinformation). For example:

- Disinformation may show signs of coordinated amplification, where a group of accounts simultaneously shares the same false content.
- Misinformation may be spread more sporadically, with no clear pattern of coordination or inauthentic behavior.

8.3 MULTI-FACTOR SCORING SYSTEM

- Assessing Multiple Factors for "Fakeness": Disinformation is rarely straightforward; it often involves a mix of subtle manipulations, falsehoods, or distortions. To accurately determine the likelihood that a piece of content is fake or manipulated, the tool should implement a multi-factor scoring system that evaluates various aspects of the content. The scoring system should use different criteria to assess the probability that the content is disinformation, assigning a score based on a combination of factors such as:
  - Source Reliability: The system should evaluate the credibility of the source that posted the content. For instance, content originating from recognized and trusted news organizations may be scored lower in terms of "fakeness" compared to content from unverified or dubious sources, such as anonymous blogs, newly created accounts, or websites with a history of spreading false information.
  - Content Accuracy and Verifiability: The system should compare the content against known facts, using databases of verified information to check for factual inaccuracies or outright falsehoods. For example, a news article claiming that a specific event occurred can be cross-referenced with fact-checked sources to determine whether the claim is verifiable.
  - Emotional Manipulation: Disinformation often uses emotionally charged language designed to provoke fear, anger, or outrage, which drives users to share the content more widely. The tool should assess whether the content uses hyperbolic or inflammatory language as a tactic to manipulate user emotions, raising the likelihood of it being disinformation.
  - Amplification by Bots: If the content has been amplified by bots or fake accounts, this should increase the "fakeness" score. Disinformation campaigns often use coordinated networks of bots to artificially boost the visibility of content, making it appear more legitimate than it actually is. The tool can analyze engagement patterns to determine whether bot networks are involved in amplifying the content.
  - Consistency with Disinformation Patterns: The system should analyze whether the content aligns with known disinformation patterns, such as previous examples of hoaxes,

Funded by the European Union

false claims, or conspiracy theories. By recognizing these patterns, the tool can detect subtle forms of disinformation that might evade more straightforward fact-checking.

- Multi-Factor Scoring for Both: The multi-factor scoring system should assess both disinformation and misinformation by using factors such as source reliability, factual accuracy, and engagement patterns. For disinformation, additional factors like bot amplification and coordinated activity can raise the content's "fakeness" score. For misinformation, the score might be lower, but the tool can still flag the content as potentially misleading, especially if it contains factual inaccuracies or emotionally manipulative language
    - Example: A post shared by a bot network that uses exaggerated language to attack a political figure would score higher for "fakeness" as disinformation. In contrast, a user sharing a factually incorrect but non-coordinated article about a health topic might receive a lower "fakeness" score, indicating misinformation rather than disinformation.
- Score Ranges and Interpretation: The tool should use a scoring system with distinct ranges that indicate the likelihood of a piece of content being fake. For example:
    - 0-25%: Likely trustworthy (little to no signs of disinformation).
    - 26-50%: Potentially misleading (some signs of manipulation or inaccuracies).
    - 51-75%: Highly questionable (clear evidence of manipulation, source issues, or emotional manipulation).
    - 76-100%: Confirmed disinformation (strong signs of coordinated amplification, factual inaccuracies, or other red flags).

8.4 TRANSPARENCY IN SCORING

- Explaining the Score: Transparency is essential for users to trust the tool's decisions. The tool must provide detailed explanations for why content was flagged as fake or misleading, breaking down the factors that contributed to the score. For example, if a news article receives a high "fakeness" score, the tool should explain that this was due to a combination of unreliable sourcing, emotionally manipulative language, and bot amplification.
    - Clear Breakdown of Indicators: Each flagged piece of content should come with a clear breakdown of the indicators that contributed to its score. For instance, a post flagged for disinformation might show that:
        - The source is unverified or has a history of spreading false claims.
        - The content contains factual inaccuracies according to cross-referenced fact-checking sources.
        - The post was retweeted or shared disproportionately by bot networks.

Funded by the European Union

- Transparency in Differentiating Disinformation and Misinformation
    - Clear Explanations to Users: The tool should provide clear explanations of whether the flagged content is more likely to be disinformation or misinformation. For disinformation, the tool can explain that the content appears to be part of a coordinated effort to deceive and that certain behaviors, such as bot activity or repeated sharing by inauthentic accounts, indicate deliberate intent. For misinformation, the tool should explain that while the content is false or misleading, it is likely being shared by individuals who believe it to be true.
        - Example: A notification to the user could explain, "This content appears to be disinformation based on its origin from unverified sources and amplification by bot accounts," or alternatively, "This post contains misinformation. While it is inaccurate, there is no evidence of coordinated efforts to spread this information intentionally."

This level of transparency not only helps users understand why the content was flagged but also allows them to make informed decisions about whether to trust or engage with it.

- Contextual Explanations: The tool should also provide contextual explanations for certain decisions, particularly in cases where the content may be nuanced or complex. For example, if the content is flagged because it uses emotionally charged language, the tool should explain why this type of language is a common tactic in disinformation campaigns. By providing users with these contextual insights, the tool helps them learn how to identify disinformation patterns independently.
- Opportunity to Dispute or Review: To ensure fairness, users should have the option to dispute the score or review the flagged content. This feature is especially important for content creators or journalists whose work may be wrongly flagged as disinformation due to algorithmic errors. Allowing users to submit evidence or request a manual review of their content ensures that the system remains fair and responsive.
    - Example: A user who writes an opinion piece that is flagged as disinformation could contest the flagging by providing sources for their claims, prompting a re-evaluation of the content.

8.5 EDUCATIONAL RESOURCES

- Empowering Users with Knowledge: One of the most effective ways to combat disinformation is through education. The tool should include a robust set of educational resources that help users understand how disinformation works and how to spot it in the future. These resources should be easy to access from within the tool and should offer users actionable insights on how to improve their media literacy.
    - Guides on Identifying Disinformation: Provide step-by-step guides or tutorials that teach users the common signs of disinformation, such as sensational headlines, unverified claims, and the use of emotional manipulation to provoke reactions. These guides could

Funded by the European Union

also explain how disinformation campaigns leverage bots and fake accounts to amplify content.

- Trusted Fact-Checking Organizations: Offer users a list of trusted fact-checking organizations that they can turn to for independent verification of questionable content. This list should be curated based on credibility, reliability, and neutrality, ensuring that users have access to accurate information when they need to verify claims.
- Interactive Tools for Self-Education: Developers can integrate interactive tools such as quizzes, exercises, or simulations that allow users to test their ability to spot fake news or disinformation. These tools could provide real-world examples of fake news articles, allowing users to apply what they've learned about disinformation detection in a controlled setting.
- Case Studies of Past Disinformation Campaigns: Provide users with case studies of past disinformation campaigns, breaking down how these campaigns were structured, how they spread, and what tactics were used to deceive people. These case studies can offer valuable lessons on how disinformation operates across different platforms, regions, and topics.

- User Education on the Differences: The tool should include educational resources that help users understand the difference between disinformation and misinformation. This can include guides or videos explaining how disinformation is often intentionally misleading and part of a broader agenda, while misinformation is shared out of a genuine misunderstanding or lack of knowledge. By providing users with this context, the tool empowers them to critically assess content before sharing it further.
  - Example: A tutorial within the tool could show real-world examples of both disinformation and misinformation, explaining the motivations behind each and teaching users how to spot the subtle differences in how they are shared and spread.
- Encouraging Critical Thinking: Beyond just flagging content, the tool should encourage users to think critically about the information they encounter. Resources that teach users how to fact-check content, identify sources, and spot manipulative tactics (such as clickbait headlines or emotionally charged language) can help prevent both disinformation and misinformation from spreading. Users should also be directed toward trusted sources and fact-checking platforms for further verification.
- Ongoing Learning and Updates: Disinformation tactics are constantly evolving, so the tool should regularly update its educational resources to reflect the latest trends and challenges. This ensures that users are always aware of new disinformation strategies and have the tools they need to protect themselves from emerging threats.
  - Example: An update to the educational resources might include information on how deepfakes or AI-generated content are being used in new disinformation campaigns, giving users the knowledge they need to recognize these emerging tactics.

Funded by the European Union

8.6 IMPLEMENTATION RECOMMENDATIONS

1. Develop a Multi-Factor Scoring Model: Implement a scoring system that assesses content based on multiple factors, including source reliability, factual accuracy, emotional manipulation, and amplification by bots. This ensures a holistic evaluation of content's likelihood of being disinformation.
2. Ensure Transparency in Flagging: Provide users with clear, detailed explanations for why content is flagged as disinformation, breaking down the specific factors that contributed to the score. Include an option for users to dispute or review flagged content.
3. Offer Comprehensive Educational Resources: Equip the tool with educational resources such as guides, fact-checking tools, and interactive modules to help users understand disinformation and how to identify it on their own.

8.7 CONCLUSION

Explaining "fakeness" is crucial to the success of AI-driven disinformation detection tools. By clearly differentiating between disinformation and misinformation, the AI tool provides users with a nuanced understanding of the "fakeness" of content. Through a multi-factor scoring system, the tool assesses content based on its origin, intent, and amplification patterns, helping users distinguish between maliciously spread falsehoods and unintentionally shared misinformation. Transparency in explaining these differences, coupled with educational resources, empowers users to make informed decisions about the content they encounter and share, ultimately reducing the spread of false information across platforms.

Funded by the European Union

9 STAKEHOLDERS' INTEGRATION

Stakeholders' integration is a critical component in the development and continuous improvement of AI tools designed to combat disinformation. By involving journalists, researchers, policymakers, and other key stakeholders, developers can ensure that the tool addresses real-world needs, adapts to emerging challenges, and reflects a wide range of perspectives. This collaborative approach allows for the tool to be regularly updated and refined, making it more effective in detecting and countering evolving disinformation tactics. Below is an expanded explanation of how stakeholder integration can be effectively incorporated into the development process.

9.1 INCORPORATING STAKEHOLDERS' FEEDBACK

- Engaging a Broad Range of Stakeholders: To ensure that the tool is useful and applicable across different sectors, developers must engage a diverse group of stakeholders throughout the development and refinement process. This includes not only developers and data scientists but also journalists, fact-checkers, researchers, policymakers, and civil society organizations. Each of these groups brings a unique perspective on the challenges posed by disinformation, and their feedback is crucial in shaping the tool's functionality.
    - Journalists and Fact-Checkers: Journalists and fact-checkers are often on the front lines of combating disinformation, so their insights into how disinformation is spread and how it is detected are invaluable. Developers should work closely with journalists to ensure that the tool can flag and analyze content that is relevant to their reporting, whether it involves disinformation campaigns around elections, health crises, or geopolitical events. By incorporating the experiences of these professionals, the tool can be fine-tuned to catch the types of disinformation they encounter in their daily work.
    - Researchers and Academics: Researchers and academics can offer important insights into disinformation patterns, emerging trends, and socio-political contexts that might not be immediately apparent to developers. By engaging with these stakeholders, developers can stay informed about the latest research on disinformation tactics, ensuring that the tool evolves in line with academic advancements.
    - Policymakers: Policymakers play a significant role in shaping legislation and guidelines around disinformation, content moderation, and online safety. Consulting with policymakers allows developers to ensure that the tool aligns with existing regulations and ethical guidelines and that it can be effectively deployed within policy frameworks to combat disinformation at scale.
- Regular Feedback Loops: Developers should implement a system of regular feedback loops to ensure that the tool stays aligned with the needs of stakeholders. This could include workshops, focus groups, or surveys where stakeholders can provide input on the tool's performance, its

Funded by the European Union

usability, and any gaps they have identified. This iterative process helps ensure that the tool is constantly being refined based on practical, real-world input from those who are directly impacted by disinformation.

- Example: After releasing a new feature for identifying disinformation on social media platforms, developers could gather feedback from journalists and fact-checkers on how effectively the feature helps them detect false content and what improvements could be made.

9.2 BETA TESTING WITH USERS

- Diverse Beta Testing Groups: In addition to gathering stakeholder feedback, it is essential to involve beta testers from various backgrounds in the development process to ensure the tool is user-centric, adaptable, and practical in diverse settings. Beta testing allows developers to see how the tool performs in real-world conditions and identify potential usability issues before it is widely deployed. By recruiting testers from a range of professions, regions, and technical backgrounds, developers can ensure that the tool works effectively for everyone—from media professionals to civil society groups and everyday users.
    - Testing with Media Professionals: Journalists, media analysts, and fact-checkers can help beta test the tool by using it to flag disinformation in news stories, social media posts, or online platforms. Their feedback on ease of use, accuracy, and real-time analysis will be crucial in fine-tuning the tool's interface and functionality.
    - Testing with Policy Experts and Governments: Government officials and policymakers can provide insights into how the tool can be used to monitor disinformation campaigns during sensitive periods such as elections. Their feedback could lead to improvements in the tool's ability to track large-scale, coordinated disinformation campaigns, particularly when linked to foreign interference or national security issues.
    - Testing with Civil Society Groups and the Public: Civil society groups and individual users can provide valuable feedback on how the tool impacts digital literacy and content moderation. By engaging a broad range of testers from different geographical and cultural backgrounds, developers can assess how well the tool performs in detecting localized disinformation that might be specific to a region or community.
- Iterative Development Based on Testing: The insights gained from beta testing should be incorporated into an iterative development process where the tool is continuously updated and improved. Developers should respond to feedback by making modifications to the tool's design, features, and functionalities, ensuring that it is as effective and user-friendly as possible.
    - Example: After a round of beta testing with journalists and researchers, developers might learn that the tool needs to provide more granular control over the types of

Funded by the European Union

disinformation it flags (e.g., political vs. health-related disinformation). The developers could then refine the tool to include filters that allow users to tailor their experience.

9.3 STAYING UPDATED ON NEW TACTICS

- Monitoring Evolving Disinformation Tactics: Disinformation tactics are constantly evolving as platforms introduce new features, as AI-generated content (e.g., deepfakes) becomes more sophisticated, and as malicious actors find new ways to manipulate the spread of information. To ensure that the tool remains effective, developers must actively monitor these emerging disinformation tactics and regularly update the tool's algorithms to recognize and counter them.
    - Tracking New Forms of Disinformation: Developers should stay informed on the latest trends in disinformation dissemination, including the use of AI-generated fake content, deepfakes, and synthetic media. This might involve collaborating with research institutions, monitoring social media platforms for new disinformation trends, or engaging with experts in digital forensics.
    - Adapting to New Platform Features: Social media platforms regularly introduce new features, such as ephemeral content (e.g., stories that disappear after 24 hours) or encrypted messaging, which disinformation campaigns may exploit. Developers need to ensure that the tool is continuously updated to adapt to these new features and to maintain the ability to detect disinformation even as platforms evolve.
- Continuous Feedback Integration: Just as new disinformation tactics are constantly emerging, stakeholder feedback should be continually integrated into the development cycle. Developers should maintain a network of experts who can provide ongoing insights into how disinformation evolves and offer real-time feedback on the tool's performance in different environments. This ensures that the tool is constantly updated and capable of addressing new challenges.
    - Example: After identifying an uptick in the use of deepfake videos in disinformation campaigns, developers might update the tool's detection algorithms to include video analysis capabilities that flag AI-generated or manipulated video content.
- Collaborating with Fact-Checkers and Experts: Developers should actively collaborate with fact-checking organizations and academic researchers who specialize in disinformation to ensure that the tool is aware of new disinformation strategies. This collaboration can provide the tool with access to databases of flagged disinformation content, expert analyses, and case studies, helping the tool stay ahead of emerging tactics.

Funded by the European Union

9.4 IMPLEMENTATION RECOMMENDATIONS

1. Establish Regular Stakeholder Engagement: Set up a process for regularly consulting journalists, researchers, policymakers, and other relevant stakeholders through workshops, focus groups, and advisory boards. This ensures that the tool evolves in line with real-world needs.
2. Design a Comprehensive Beta Testing Program: Recruit a diverse range of beta testers from different regions, sectors, and backgrounds to rigorously test the tool before it is released more broadly. Incorporate their feedback into iterative improvements.
3. Monitor and Adapt to New Disinformation Tactics: Keep the tool's detection algorithms updated to recognize new disinformation strategies, including deepfakes, AI-generated content, and evolving platform features. This ensures that the tool remains relevant and effective in a constantly changing digital landscape.

9.5 CONCLUSION

Incorporating stakeholder feedback, engaging diverse beta testers, and staying updated on emerging disinformation tactics are essential to the ongoing success of any AI-driven disinformation detection tool. By consulting key stakeholders and continuously integrating feedback, developers can ensure that the tool remains relevant, effective, and responsive to the evolving digital landscape. This collaborative approach enhances the tool's ability to meet the needs of various sectors, while also ensuring that it adapts to new challenges in disinformation detection and mitigation.

Funded by the European Union

10 CONTINUOUS IMPROVEMENT

To maintain relevance, accuracy, and effectiveness in a rapidly evolving digital landscape, AI-driven disinformation detection tools must be built with mechanisms for continuous improvement. This involves regular updates to algorithms, integration of user feedback, and ongoing collaboration with disinformation experts and researchers. The goal is to ensure that the tool remains responsive to new challenges, including emerging disinformation tactics and evolving platform features. Below is an in-depth look at how continuous improvement can be incorporated into the development and operation of such tools.

10.1 REGULAR ALGORITHM UPDATES

- Adapting to New Disinformation Trends: Disinformation strategies evolve quickly, as bad actors find new ways to manipulate information and exploit vulnerabilities in platforms. To keep pace, the tool's algorithms must be updated regularly to reflect the latest disinformation trends. These updates should focus on improving the tool's ability to detect emerging forms of disinformation, such as the use of deepfakes, AI-generated text, or sophisticated bot networks designed to mimic human behavior.
    - Tracking New Techniques: Developers should actively track new techniques used to spread disinformation, whether through social media platforms, messaging apps, or news aggregation sites. This could involve monitoring foreign interference campaigns, political disinformation, or even the use of manipulated imagery to deceive viewers. As these techniques evolve, developers need to regularly refine the algorithms to recognize the latest patterns and methods.
    - Expanding Data Sources: The tool's effectiveness is heavily reliant on the datasets used to train its algorithms. As disinformation tactics evolve, it is important to expand the sources of data used to train the tool. This could include incorporating new datasets from different regions, updated lists of disinformation sites, and data from fact-checking organizations. Regularly updating the datasets ensures that the tool can detect a wider range of disinformation from various contexts.
    - Improving Detection Accuracy: Algorithm updates should not only focus on new trends but also on improving the overall accuracy and precision of detection. Regular reviews of the tool's performance, including identifying any false positives or false negatives, should inform the development of more precise detection techniques. These updates may involve refining natural language processing (NLP) models, enhancing bot-detection algorithms, or incorporating better image and video analysis tools.
    - Example: If a new form of AI-generated content becomes prevalent in disinformation campaigns, developers might update the tool's algorithms to specifically target and

Funded by the European Union

analyze this content. This could involve training the model on a new dataset of AI-generated deepfakes to enhance detection capabilities.

10.2 USER FEEDBACK MECHANISMS

- Encouraging User Engagement: Continuous improvement requires ongoing input from users, who can provide valuable insights into how the tool performs in real-world scenarios. Implementing robust user feedback mechanisms is essential for collecting suggestions, identifying pain points, and improving the tool's functionality based on user experience. This could take the form of in-tool feedback buttons, surveys, or customer support channels where users can share their experiences and report issues.
    - Customizing Feedback Options: Users should have multiple ways to provide feedback depending on their level of expertise or the issues they encounter. For example, journalists and researchers might submit more detailed feedback on how well the tool flags disinformation in complex news stories, while general users might highlight user interface issues or false flagging of benign content. By offering both structured surveys and open-ended feedback forms, developers can gather a wide range of insights.
    - Feedback for Continuous Refinement: Once feedback is gathered, it should be incorporated into regular development sprints to ensure that the tool evolves based on user needs. Prioritizing frequently mentioned issues—such as improving the clarity of flagged content explanations or refining the flagging accuracy—can lead to more user-centric enhancements.
    - Example: A group of beta testers might report that the tool is flagging too many benign posts as disinformation, suggesting a need to adjust the sensitivity of the algorithms. Developers could use this feedback to fine-tune the tool's detection thresholds.
- Providing Feedback Loops: Users who submit feedback should be kept informed about how their suggestions have been addressed. Offering transparency in the feedback process by communicating when updates or improvements have been made based on user input helps build trust and engagement. This could involve sending update notifications to users or publishing release notes that highlight changes made in response to user feedback.

Funded by the European Union

10.3 COLLABORATION WITH EXPERTS

- Leveraging Expert Knowledge: Collaborating with disinformation experts, such as researchers, academics, and organizations focused on media integrity, is crucial for ensuring that the tool stays current with cutting-edge research and emerging threats. Experts in areas like digital forensics, AI-generated content, and propaganda tactics can provide valuable insights that help developers fine-tune the tool's detection mechanisms and anticipate future challenges.
    - Engaging with Research Communities: Developers should maintain close relationships with academic and research communities that focus on disinformation, misinformation, and online manipulation. Attending conferences, participating in research collaborations, and contributing to academic papers can help developers stay updated on new disinformation strategies, as well as advancements in detection methods.
    - Collaborating with Fact-Checking Organizations: Fact-checking organizations are often at the forefront of identifying and debunking disinformation. Partnering with these organizations allows developers to access real-time data on ongoing disinformation campaigns, ensuring that the tool is updated with the most current examples of false content. This collaboration can also help developers improve the tool's ability to distinguish between credible and non-credible sources, based on the expertise of fact-checkers.
    - Government and Policy Engagement: Developers should also collaborate with government agencies and policymakers who are tasked with combating disinformation at a national or international level. These stakeholders can provide valuable insights into foreign interference campaigns, election-related disinformation, and other politically motivated disinformation efforts. By collaborating with these groups, developers can ensure that the tool is equipped to handle the most pressing disinformation threats.
- Incorporating the Latest Research: Disinformation detection is an area of active research, with new methods and approaches being developed regularly. Developers should regularly review academic literature, participate in industry workshops, and engage in knowledge exchange with experts to stay informed about the latest breakthroughs. By continuously integrating findings from the field, developers can ensure that the tool remains at the forefront of disinformation detection technology.
    - Example: After collaborating with a group of academics researching deepfakes, developers could implement new detection techniques that analyze subtle artifacts in AI-generated videos, improving the tool's ability to flag manipulated content.

Funded by the European Union

10.4 IMPLEMENTATION RECOMMENDATIONS

1. Schedule Regular Algorithm Updates: Implement a process for regularly updating the tool's algorithms and datasets to reflect new disinformation trends, tactics, and platform features. This includes tracking emerging tactics like deepfakes, AI-generated text, and evolving social media strategies.
2. Create User Feedback Channels: Establish multiple feedback channels that allow users to provide input on the tool's performance, report issues, and suggest improvements. Regularly review and act on this feedback, ensuring that user insights shape future development.
3. Collaborate with Experts and Researchers: Maintain ongoing collaboration with disinformation experts, fact-checking organizations, and academic researchers to ensure that the tool incorporates the latest research and stays current with emerging threats. Engage in industry workshops and contribute to the field of disinformation detection.

10.5 CONCLUSION

Continuous improvement is essential for the success and longevity of AI-driven disinformation detection tools to trigger resilience mechanisms. By regularly updating algorithms, integrating user feedback, and collaborating with disinformation experts, developers can ensure that the tool remains accurate, effective, and responsive to new challenges. These efforts are critical to keeping the tool relevant in the face of constantly evolving disinformation tactics and ensuring that users have the most up-to-date protection against misinformation and manipulation.

Funded by the European Union

11 CONCLUSION

AI tools for disinformation detection are crucial in the fight against the spread of false information online, but their effectiveness depends on a continuous commitment to improvement and adaptability to trigger resilience mechanisms. The landscape of disinformation is constantly changing, with new tactics and technologies emerging to circumvent detection. As such, developers must prioritize regular updates to algorithms, the integration of user feedback, and collaboration with experts in the field to keep these tools current and effective.

Furthermore, the ethical considerations of data privacy, transparency, and bias avoidance are essential to ensuring that these tools operate fairly and responsibly. By engaging a wide range of stakeholders, including journalists, researchers, and policymakers, developers can ensure that the tools remain relevant, responsive, and tailored to real-world needs.

In conclusion, AI disinformation detection tools must be designed with flexibility and continuous improvement at their core. Through regular updates, ethical safeguards, and ongoing collaboration, these tools can help maintain the integrity of public discourse and prevent the damaging effects of false information in an increasingly digital world.

Funded by the European Union

REFERENCES

Anzalone, R. (2019, November 1). *AI-altered video is a threat to society—How do we stop the harm deepfakes can cause?* Forbes. Retrieved December 12, 2024, from https://www.forbes.com/sites/robertanzalone/2019/11/01/ai-altered-video-is-a-threat-to-society--how-do-we-stop-the-harm-deepfakes-can-cause/

Arif, M., Tonja, A. L., Ameer, I., Kolesnikova, O., Gelbukh, A., Sidorov, G., & Meque, A. G. M. (2022). *CIC at CheckThat! 2022: Multi-class and cross-lingual fake news detection*. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum* (Vol. 3180). CEUR-WS.org. Retrieved December 12, 2024, from https://ceur-ws.org/Vol-3180/paper-33.pdf

Arounda Agency. (n.d.). *Accessibility in UX: Best practices and key principles*. Retrieved December 12, 2024, from https://arounda.agency/blog/accessibility-in-ux-best-practices-and-key-principles

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020, July 28). *Exposure to social engagement metrics increases vulnerability to misinformation*. HKS Misinformation Review. Retrieved December 12, 2024, from https://misinforeview.hks.harvard.edu/article/exposure-to-social-engagement-metrics-increases-vulnerability-to-misinformation/

Bateman, J., & Jackson, D. (2024, January 31). *Countering disinformation effectively: An evidence-based policy guide*. Carnegie Endowment for International Peace. Retrieved December 12, 2024, from https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide?lang=en

Belissent, J. (2024, February 6). *5 steps to data diversity: More diverse data makes for smarter AI*. Snowflake. Retrieved December 12, 2024, from https://www.snowflake.com/en/blog/five-steps-data-diversity-for-smarter-ai-models/

Canadian Centre for Cyber Security. (2024, May). How to identify misinformation, disinformation, and malinformation (ITSAP.00.300). Retrieved December 12, 2024, from https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300

Carr, R., & Köhler, P. (2024, October). *AI-pocalypse Now? Disinformation, AI, and the Super Election Year*. Munich Security Conference. Analysis 4/2024. https://securityconference.org/en/publications/analyses/ai-pocalypse-disinformation-super-election-year/

CEN and CENELEC. (2024, September). *How to be gender-responsive in standardization*. Retrieved December 12, 2024, from https://www.cencenelec.eu/media/CEN-CENELEC/News/Publications/2024/genderresponsivestandardization_access.pdf

Chipeta, C. (2024, November 18). *Open source intelligence (OSINT): Top tools and techniques*. UpGuard. Retrieved December 12, 2024, from https://www.upguard.com/blog/open-source-intelligence

Commons Librarian. (n.d.). Disinformation vs Misinformation: Definitions & Types. The Commons Social Change Library. Retrieved December 12, 2024, from https://commonslibrary.org/disinformation-vs-misinformation-definitions-types/

Concepta Tech. (2023, August 14). *How to define stakeholders for your software development project*. Retrieved December 12, 2024, from https://www.conceptatech.com/blog/how-to-define-stakeholders-for-your-software-development-project

Contentsquare. (n.d.). *User feedback examples: How to collect and leverage feedback for better UX*. Retrieved December 12, 2024, from https://contentsquare.com/guides/user-feedback/examples/

Corporate English Solutions. (2023, October 30). *Minimising AI bias: Best practices for organisations*. British Council. Retrieved December 12, 2024, from https://corporate.britishcouncil.org/insights/minimising-ai-bias-best-practices-organisations

Crest Infotech. (n.d.). *Ethical considerations in mobile app development*. Retrieved December 12, 2024, from https://www.crestinfotech.com/ethical-considerations-in-mobile-app-development/

Cser, T. (2022, December 19). *The importance of diversity in software testing teams: The benefits of diversity in software testing teams that include people with various backgrounds, experiences, perspectives, and skill sets*. Retrieved December 12, 2024, from https://www.functionize.com/blog/importance-of-diversity-in-software-testing-teams

Data Security Council of India. (n.d.). *Unmasking the false: Advanced tools and techniques for deepfake detection*. Cybersecurity Centre of Excellence. Retrieved December 12, 2024, from https://ccoe.dsci.in/blog/deepfake-detection

Dilmegani, C. (2024, November 18). *Top 12 AI ethics dilemmas: Real-life examples & tips to mitigate*. AIMultiple. Retrieved December 12, 2024, from https://research.aimultiple.com/ai-ethics/

Do, Q. N. T., & Gaspers, J. (2019, October 28). *Improving cross-lingual transfer learning by filtering training data*. Amazon Science. Retrieved December 12, 2024, from https://www.amazon.science/blog/improving-cross-lingual-transfer-learning-by-filtering-training-data

Document360. (2022, May 10). *Create multilingual documentation: Steps, benefits, and best practices*. Retrieved December 12, 2024, from https://document360.com/blog/multilingual-documentation/

European Commission. (2022, June 16). *The 2022 Code of Practice on Disinformation*. Shaping Europe's Digital Future. Retrieved December 12, 2024, from https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

European Commission. (2022, October 11). *Guidelines for teachers and educators on tackling disinformation and promoting digital literacy through education and training*. Retrieved December 12, 2024, from https://education.ec.europa.eu/news/guidelines-for-teachers-and-educators-on-tackling-disinformation-and-promoting-digital-literacy-through-education-and-training

European Data Protection Supervisor. (2024, November 30). *Fake news detection*. Retrieved December 12, 2024, from https://www.edps.europa.eu/press-publications/publications/techsonar/fake-news-detection_en

Fairman, G. (2021, October 11). *Six localization tips for multilingual customer support*. Bureau Works. Retrieved December 12, 2024, from https://www.bureauworks.com/blog/6-localization-tips-for-multilingual-customer-support-fc

Feast, J. (2023, October 10). *How to identify and mitigate gender bias in AI*. Cogito. Retrieved December 12, 2024, from https://cogitocorp.com/blog/how-to-identify-and-mitigate-gender-bias-in-ai/

GeeksforGeeks. (n.d.). *Software engineering: User interface design*. Retrieved December 12, 2024, from https://www.geeksforgeeks.org/software-engineering-user-interface-design/

Funded by the European Union

Government Communication Service. (2021, November). *RESIST 2 counter-disinformation toolkit*. Retrieved December 12, 2024, from https://gcs.civilservice.gov.uk/publications/resist-2-counter-disinformation-toolkit/

Google PAIR. (n.d.). *Explainability and trust*. Retrieved December 12, 2024, from https://pair.withgoogle.com/chapter/explainability-trust/

Grennan, L., Kremer, A., Singla, A., & Zipparo, P. (2022, September 29). *Why businesses need explainable AI—and how to deliver it*. McKinsey & Company. Retrieved December 12, 2024, from https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it

Hawke, J. (2023). *AI Assistant Mediators: New Tools for Better Conversations*. MIT Solve. Retrieved December 12, 2024, from https://solve.mit.edu/challenges/learning-for-civic-action-challenge/solutions/72310

Insikt Group. (2024, September 24). *Targets, objectives, and emerging tactics of political deepfakes*. Recorded Future. Retrieved December 12, 2024, from https://www.recordedfuture.com/research/targets-objectives-emerging-tactics-political-deepfakes

Kramer, N. (2024, May 15). *Integrating user feedback in software development: 10 strategies*. daily.dev. Retrieved December 12, 2024, from https://daily.dev/blog/integrating-user-feedback-in-software-development-10-strategies

Komendantova, N., Ekenberg, L., Svahn, M., & others. (2021). A value-driven approach to addressing misinformation in social media. *Humanities and Social Sciences Communications, 8*(33). https://doi.org/10.1057/s41599-020-00702-9

Kuntur, S., Wróblewska, A., Paprzycki, M., & Ganzha, M. (2024).*Fake news detection: It's all in the data!* arXiv. https://arxiv.org/abs/2407.02122

Li, C., & Callegari, A. (2024, June). *How AI is being used to combat online misinformation and disinformation*. World Economic Forum. Retrieved December 12, 2024, from https://www.weforum.org/stories/2024/06/ai-combat-online-misinformation-disinformation/

LinkedIn Community. (n.d.). *Content strategy: What are the most effective ways to measure content virality?* Powered by AI and the LinkedIn community. Retrieved December 12, 2024, from https://www.linkedin.com/advice/1/what-most-effective-ways-measure-content-virality-edzzc

LinkedIn. (n.d.). How can you use stakeholder feedback to refine your technology? Retrieved December 12, 2024, from https://www.linkedin.com/advice/0/how-can-you-use-stakeholder-feedback-refine-your-technology-quvqe

Mailchimp. (n.d.). *AI transparency: Building trust in AI*. Retrieved December 12, 2024, from https://mailchimp.com/resources/ai-transparency/

Manasi, A., Panchanadeswaran, S., & Sours, E. (2023, March 17). *Addressing gender bias to achieve ethical AI*. IPI Global Observatory. Retrieved December 12, 2024, from https://theglobalobservatory.org/2023/03/gender-bias-ethical-artificial-intelligence/

Meta. (n.d.). *Community standards: Inauthentic behavior*. Transparency Center. Retrieved December 12, 2024, from https://transparency.meta.com/en-gb/policies/community-standards/inauthentic-behavior/

Novoselova, O. V. (2021, January 26). *Fake news & stakeholder management skills*. Leadership & Flow. Retrieved December 12, 2024, from https://flowleadership.org/fakenews/

OECD (2024), *The OECD Reinforcing Democracy Initiative: Monitoring Report – Assessing Progress and Charting the Way Forward*, OECD Public Governance Reviews, OECD Publishing, Paris, https://doi.org/10.1787/9543bcfb-en.

Omdena. (2023, December 13). *The ethical role of AI in media: Combating misinformation*. Retrieved December 12, 2024, from https://www.omdena.com/blog/the-ethical-role-of-ai-in-media-combating-misformation

POEditor. (n.d.). *Software localization: What it is and why it matters*. Retrieved December 12, 2024, from https://poeditor.com/blog/software-localization

Phyllo. (2024, July 30). *Social media API: Guide on top APIs for developers*. Retrieved December 12, 2024, from https://www.getphyllo.com/post/social-media-api-guide-on-top-apis-for-developers

Ramanathan, A. (2024, November 15). *Multilingual customer support: 7 strategies to attract global customers*. DevRev. Retrieved December 12, 2024, from https://devrev.ai/blog/multilingual-customer-support

Rouse, M. (n.d.). *AI transparency: What is it and why do we need it?* TechTarget. Retrieved December 12, 2024, from https://www.techtarget.com/searchcio/tip/AI-transparency-What-is-it-and-why-do-we-need-it

Sahota, N. (2023, September 19). *Technological solutions for deepfake detection during elections*. LinkedIn. Retrieved December 12, 2024, from https://www.linkedin.com/pulse/technological-solutions-deepfake-detection-during-elections-sahota/

SAP. (n.d.). *What is AI ethics?* Retrieved December 12, 2024, from https://www.sap.com/resources/what-is-ai-ethics

Shivani Nayuni, I. E. (2024, November 5). *AI-driven tools and techniques for fake news detection*. Insights2TechInfo. Retrieved December 12, 2024, from https://insights2techinfo.com/ai-driven-tools-and-techniques-for-fake-news-detection/

Søe, S.O. (2018), "Algorithmic detection of misinformation and disinformation: Gricean perspectives", *Journal of Documentation*, Vol. 74 No. 2, pp. 309-332. https://doi.org/10.1108/JD-05-2017-0075

Split. (2024, June 14). *Enhancing product development with user feedback loops*. Retrieved December 12, 2024, from https://www.split.io/blog/enhancing-product-development-with-user-feedback-loops/

Stempeck, M. (2024, February 29). *9 partners governments can team up with to counter disinformation*. Democracy Technologies. Retrieved December 12, 2024, from https://democracy-technologies.org/disinformation/9-partners-to-counter-disinformation/

Sun, H. (2023). *Regulating algorithmic disinformation*. *Columbia Journal of Law & the Arts, 46*(3), 367-397. https://journals.library.columbia.edu/index.php/lawandarts/article/download/11237/5582/28927

Taylor & Francis. (n.d.). Misinformation vs. Disinformation. Retrieved December 12, 2024, from https://insights.taylorandfrancis.com/social-justice/misinformation-vs-disinformation/

Tsai, L. L., Pentland, A., Braley, A., Chen, N., Enríquez, J. R., & Reuel, A. (2024, March). *Generative AI for pro-democracy platforms*. An MIT exploration of generative AI. https://doi.org/10.21428/e4baedd9.5aaf489a

UXPin. (n.d.). *User-friendly: What does it mean and how to apply it?* Retrieved December 12, 2024, from https://www.uxpin.com/studio/blog/user-friendly-what-does-it-mean-and-how-to-apply-it/

Vaish, T. (2021, October 11). *How to manage to stay on top with social media algorithms and trends*. One Nought One. Retrieved December 12, 2024, from https://www.onenought.one/post/how-to-manage-to-stay-on-top-with-social-media-algorithms-and-trends

Volodina, K. (2024, June 3). *Social media engagement: 11 key strategies on how to increase it*. Socialinsider. Retrieved December 12, 2024, from https://www.socialinsider.io/blog/social-media-engagement/

Vorecol. (n.d.). *Best practices for integrating user feedback into software development processes*. Retrieved December 12, 2024, from https://vorecol.com/blogs/blog-best-practices-for-integrating-user-feedback-into-software-development-processes-172813

Vorecol. (n.d.). *Cultural sensitivity features in software: What multicultural teams really need*. Retrieved December 12, 2024, from https://vorecol.com/blogs/blog-cultural-sensitivity-features-in-software-what-multicultural-teams-really-need-161477

Winyama. (2023, March 19). *Exploring the role of AI in cultural sensitivity and content diversity*. Retrieved December 12, 2024, from https://www.winyama.com.au/news-room/exploring-ai-cultural-sensitivity-content-diversity

World Wide Web Consortium (W3C). (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. Retrieved December 12, 2024, from https://www.w3.org/TR/WCAG21/

Funded by the European Union

Funded by
the European Union

ANNEX 1: GUIDELINES FOR THE BETA TESTING MEETINGS

The AI4DEBUNK project WP12 foresees the set up of a beta testing group to support and test the tool's development. The guidelines for the beta testing meetings are presented below:

1. Objective

To evaluate fake news and responses (starting year 1) and the effectiveness of solution the tool designed as part of the AI4DEBUNK project (starting year 2) to detect and combat disinformation.

2. Purpose

The role of a beta tester is to provide feedback on the AI4DEBUNK solutions to help improve it, ensuring it meets users' needs and effectively combats disinformation.

3. Roles

Beta testers' role is to test methods, including the tool, simulating real-world conditions. They follow the methodologies as, as instructed, by the moderator.

4. Expectations:

a) Test how fake news affects public perception, exploring its influence on opinions and its broader impact in various contexts.

b) Evaluate the responses generated by the AI4DEBUNK tool to determine if it accurately detects and reacts to disinformation.

c) Test the tool's functionalities in various scenarios to ensure it performs well across different conditions, platforms, and content types.

d) Participate in feedback sessions or focus groups to share insights and help refine the tool's design and capabilities.

e) Keep all information related to the tool and the beta testing process confidential, refraining from sharing any details with unauthorized parties.

5. Testing Process

a) Follow the provided instructions that outline specific tasks and scenarios to be tested. If clarification is needed, seek guidance from the testing team.

b) Complete the tasks assigned during testing, such as analysing fake news examples, interacting with the tool in various ways, and testing its accuracy.

c) Evaluate the tool's effectiveness based on specific criteria, including its accuracy, response time, and reliability in detecting disinformation.

d) Offer detailed feedback on the testing experience, noting any bugs, glitches, or malfunctions, and suggesting improvements.

e) Report on the tool's usability, focusing on ease of use, interface design, and overall user experience. Highlight any areas that need improvement.

Funded by the European Union

f) Test the tool's accuracy in distinguishing between real and false information, providing feedback on its reliability.

g) Suggest potential enhancements to the tool's functionality, accuracy, or user experience based on the testing process.

6. Communication and Reporting

a) A feedback form will be provided to beta testers for reporting their experiences and observations during the testing process. This form will include sections for documenting their interactions with the tool, noting any technical issues, and providing general feedback on usability, functionality, and overall effectiveness in detecting disinformation.

b) All beta testers must understand the importance of maintaining confidentiality regarding the beta testing process. Information from the feedback form should not be shared with anyone outside the designated channels or with unauthorised parties.

7. Compensation

Beta testers should be made known right before their enrolment on the compensation that they will receive during beta testing period.

8. Code of conduct

This depending on the format of the beta tester setting (group or individual)

a) Beta testers should conduct themselves professionally during all interactions with the testing team and other beta testers.

b) Treat fellow beta testers and testing staff with respect, adhering to guidelines presented during the group discussions or focus groups.

(Detailed guidelines to be developed)

9. Selection of Beta Testers

a) When selecting beta testers, the project will ensure their anonymity is maintained to protect their identity and encourage unbiased feedback.

b) Care will be taken to understand that they have not engaged in hate speech, racism, or any form of discriminatory behaviour. In addition, the candidate beta tester should not have any criminal record

c) The project will implement basic background checks to screen for any such incidents that could make them ineligible.

d) All beta testers will be required to undergo a basic interview and also sign a consent form which will detail the above aspects before being able to join the programme.

10. Ethics and Conduct

The tool will be a developing prototype during the beta testing phase – hence no actions should be undertaken compromising its integrity.

Funded by the European Union

ANNEX II. MEETING OF THE BETA TESTING GROUP 19.06.2024

Brussels
Number of beta testers 4
Number of facilitators: 2: Pascaline Gaborit and Joen Martinsen Pilot4dev
Interests of the participants:

- Interests are diverse. Some of them have worked on GDPR privacy…Beta testeur 1 has an interest in the Western Balkans, and Elise is interested in disinformation in the aftermath of EU elections, but everyone is strictly interested as citizens and has no professional experience regarding fake news and disinformation.

Case study 1: Wind Turbines
Wind turbines after 2 years! The erosion releases fine glass fiber and epoxy resin particles, which are respirable and carcinogenic.

The beta testers highlighted the fact that you need to use fact checkers and can only rely on trustworthy media. Tiktok spreads a lot of fake news…beta testeur 2 particularly points out that he doesn't find TikTok a reliable source for any information, and no claims should be trusted from this platform. When it is revealed to them that this was first written in an article, that was then shared on TikTok.

Beta testeur 2 explained that the information of the glass fiber and epoxy resin is actually true, but the concentration is not sufficient to cause cancer. These types of "campaigns" against wind turbines are creating suspicions similar to information about nuclear energy and solar panels. Because of Beta testeur 3's expertise, the person knew that the content in the tiktok actually mostly is true regarding the erosion of wind turbines releasing glass fiber and epoxy, but the claim of this causing cancer and carcinogenic is inaccurate, so he assessed what was fake about this case study very accurately.

Beta testeur 1 pointed out that the language used on the tiktok is emotional regarding the exclamation marks…It seems that the authors address themselves to an already angry audience who are already against wind turbines.

Case study 2: WHO and Agriculture

Funded by the European Union

The World Health organization (WHO) has now called on governments around the world to ban agriculture to save the planet from what is now being called "global warming".

A new statement from WHO Director-General Dr. Tedros Adhanom Ghebreyesus has revealed that people could now secretly plan to starve people to death, albeit as part of a "necessary" reduction in the world population.

The WHO now seems to want to push this forward under the guise of fighting "climate change". According to the WHO, banning meat and dairy products is the first step in achieving its "radical goal".

It was spread on twitter, FB, Tiktok…The objective is to create confusion, but also make up a 'conspiracy theory ' or 'secret cabal'. It may target farmers who are already in protest against EUs climate policies on agriculture, which farmers have protested against across Europe. The events are presented with biases…May also discredit institutions such as WHO and create confusion between 'real' and 'unreal' information…There the beta testers think that the fake is 'too obvious'. Some however marked that the site looked more "convincing" with the dark blue theme of the website and looks more professional, which would make the news-website "Unsere MittelEuropa" convincing to some people. Beta testeur 2 noted that the only topic that are highlighted on the tag-line was "Covid-19".

Case Study 3: Pregnant women sent to the front line in Ukraine
https://vm.tiktok.com/ZGeqJYWPK/
Ukraine sends pregnant women to war
What you are about to see now is a pregnant woman surrendering to the Russian army. Ukrainian Lawmakers claim that pregnancy is not a reason to delay military service.

It is reported that Ukraine has urgently allocated 50,000 sets of female solider uniforms to support those who are about to go to the front lines. Although woman normally play auxiliary roles in war, Ukraine's predicament has reached the point where they have no choice but to send female soldiers to the front lines. However, their combat effectiveness is questionable, as most of them have never experienced combat and may show cowardice or even become captives in real battles. What's even more tragic is that the Ukrainian authorities cannot provide proper personal equipment for female soldiers. Possibly due to insufficient resources in the country.

A lot of the beta-testers had many facial expressions while the case was read out loud. Many like, like Beta testeur 2 showed signs of finding the case ludacris. Again TikTok brought up as a Beta testeur 4 said that it would not be spread on Russian media, and thought the content too far (Tone indicates that he finds the content very far-fetched, maybe surprised that someone would even believe this). However, Beta testeur 1said that similar news could be published on Serbian tabloids, that especially the title of the TikTok could have been seen on Serbian pro-Russian news outlets. The objective is that readers become schizophrenic. The authors also play with emotions and bias, something that Beta testeur 2 and Beta testeur 4 both noted.

Case Study 4: Land corridors
Revealed - NATO plan to get US troops to the front line to fight RUSSIA: Alliance prepares for rapid deployment of American soldiers amid fears Moscow is plotting major war with Europe
NATO is drawing up plans to send American troops to the frontlines of Europe in the event of an all-out conflict with Russia, it has been revealed.

New 'land corridors' are being carved out to quickly funnel soldiers through central Europe without local bureaucratic impediments, allowing NATO forces to pounce in an instant should Putin's devastating war in Ukraine move further west.

The plans are said to include contingencies in case of Russian bombardment, letting troops sweep into the Balkans via corridors in Italy, Greece and Turkey, or towards Russia's northern border via Scandinavia, officials told The Telegraph.

Published by 'The telegraph' in the UK and spread on social media. This is 'Fear mongering', unrealistic…Try to influence the support to NATO and Ukraine. Ivans initial reaction was that "i don't know if this is true, but it should be true". Didn't detect anything problematic about this information compared to the other cases. Beta testeur 3 also expressed similar sentiments about NATO preparations against Russia isn't necessarily bad, but maybe even good. Beta testeur 2 and Beta testeur 4 were surprised that the case was from the Telegraph, a UK newspaper, which to them sticks out as more trustworthy than the previous cases. When it was revealed that most of the information in the article could be true regarding the land corridors themselves (No fact-checkers have debunked the information, and previous land corridors have officially been planned. The beta testers were therefore not surprised, because they had recognized them as more realistic than the others.

Deep Fakes: can be obvious, but sometimes not and can lead to manipulation. They contribute to discredit. Is it going to lead to scams in the future (fake voices…). Beta testeur 3 marks that she is mostly worried about crime and scams related to these technologies, that it would be hard to present "proof" and evidence of a crime in a court when anything could be manipulated, and we need tools to counter this uncertainty and methods to assess what could still be considered real evidence.

Solutions and Tools

Should Deep Fakes with people be forbidden without their consent? Should Titktok be forbidden? (eg. Fake video of the niece of Marine le Pen). Or should we add a banner 'Exposure to Tiktok can damage your mental health'?. Could a firewall be created for Europe? Beta testeur 1 said that a ban is tricky…Freedom of Speech will be confronted by regulation and limits. He expresses that we shouldn't become like "our enemy", referring to China, and this highly regulated "Orwellian society" of truths, and that we shouldn't develop in similar patterns, is his argument against such a ban. Previously he has expressed that TikTok "definitely" could cause health issues, without elaborating on this.

An innovative tool would be able to debunk the authors while respecting privacy. (farm trolls or citizens)…Blockchain could be a system to control disinformation. How is it possible to have so much fake news while the registration to the platforms requires more and more checks about identity (credit cards, IDs, phone numbers)?

Distrust is important in former socialist countries. It takes time…Transparency is one of the options.
A good tool should increase trust in regular media and increase media literacy. Independent analysts could be corrupted, Beta testeur 3 marks that even fact-checkers that are not representing authorities, could still be bought, or manipulated and everyone needs financing. So it will continue to be a challenge to know where fact-checkers get their money from. Social media platforms strategies are insufficient as they do not fit their business models (and the process is based on image/news checkers sometimes exposed to violent pictures). There is also misinformation on Wikipedia
A tool could detect emotional language. The linguistic approach is also a challenge, as human language can make it hard for AI to determine if something is fake.

Funded by the European Union

Review Sheet of Deliverable/ Milestone Report
D12.2 Resilience mechanisms triggered by the tool

Editor(s):	Pascaline Gaborit, Vishnu Rao, Joen Martinsen
Responsible Partner:	Pilot4dev
Status-Version:	Draft / Final - v0.2
Date:	09/12/2024
Distribution level (CO, PU):	PU
Reviewer (Name/Organization)	Matei Mancas (UMONS)
Review date	19/12/2024

Disclaimer: This assessment reflects only the author's views and the European Commission is not responsible for any use that may be made of the information contained therein"

Funded by the European Union

Mark with X the corresponding column:

Y= yes	N= no	N = not applicable

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
FORMAT: Does the document … ?				
…include editors, deliverable name, version number, dissemination level, date, and status?	X			
…contain a license (in case of public deliverables)?	X			Copyright
…include the names of contributors and reviewers?	X			
….has a version table consistent with the document's revision?	X			
… contain an updated table of contents?	X			
… contain a list of figures consistent with the document's content?	X			
… contain a list of tables consistent with the document's content?	X			
… contain a list of terms and abbreviations?	X			
… contain an Executive Summary?	X			
… contain a Conclusions section?	X			
… contain a List of References (Bibliography) in the adequate format, if relevant?	X			
… use the fonts and sections defined in the official template?	X			
… use correct spelling and grammar?	X			
… conform to length guidelines (50 pages maximum (plus Executive Summary and annexes)	X			
… conform to guidelines regarding Annexes (inclusion of complementary information)	X			
… present consistency along the whole document in terms of English quality/style? (to avoid accidental usage of copy&paste text)	X			
About the content…				
Is the deliverable content correctly written?	X			
Is the overall style of the deliverable correctly organized and presented in a logical order?	X			
Is the Executive Summary self-contained, following the guidelines and does it include the main conclusions of the document?	X			
Is the body of the deliverable (technique, methodology results, discussion) well enough explained?	X			
Are the contents of the document treated with the required depth?	X			

Funded by the European Union

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
Does the document need additional sections to be considered complete?		X		
Are there any sections in the document that should be removed?		X		
Are all references in the document included in the references list?	X			
Have you noticed any text in the document not well referenced? (copy and paste of text/picture without including the reference in the reference list)		X		
SOCIAL and TECHNICAL RESEARCH WPs (WP4, 5, 12, 13, 14)				
Is the deliverable sufficiently innovative?	X			
Does the document present technical soundness and its methods are correctly explained?	X			Some implementation requirements would be difficult to implement but most of them should be feasible.
What do you think is the strongest aspect of the deliverable?				It treats a very wide range of aspects and it is thus a very good basis to extract feasible implementation recommendations.
What do you think is the weakest aspect of the deliverable?				The references are not included in the text, only at the end.
Please perform a brief evaluation and/or validation of the results, if applicable.				The report is valuable as it gathers a lot of interesting propositions. Some of them might be implemented during the project, others might be implemented in future projects and some might be very difficult to implement in practice. It might be interesting to get feedback from technical partners to rate the technical feasibility of the different propositions but this is probably not relevant here.
AI AND TECNOLOGICAL WPS (WP6 – WP11)				
Does the document present technical soundness and the methods are correctly explained?				
What do you think is the strongest aspect of the deliverable?				
What do you think is the weakest aspect of the deliverable?				

Funded by the European Union

ELEMENT TO REVIEW	Y	N	NA	COMMENTS
Please perform a brief evaluation and/or validation of the results, if applicable.				
DISSEMINATION AND EXPLOITATION WPs (WP15 – WP17)				
Does the document present a consistent outreach and exploitation strategy?				
Are the methods and means correctly explained?				
What do you think is the strongest aspect of the deliverable?				
What do you think is the weakest aspect of the deliverable?				
Please perform a brief evaluation and/or validation of the results, if applicable.				

SUGGESTED IMPROVEMENTS

PAGE	SECTION	SUGGESTED IMPROVEMENT

CONCLUSION

Mark with X the corresponding line.

X	Document accepted, no changes required.
	Document accepted, changes required.
	Document not accepted, it must be reviewed after changes are implemented.

Please rank this document globally on a scale of 1-5 (1 = poor, 5= excellent) – using a half point scale. Mark with X the corresponding grade.

Document grade	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**4.5**	**5**
							X		

Funded by the European Union